Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

MARIANA DE CASTRO PASQUALINI

COMPARAÇÃO DO DESEMPENHO DE MODELOS PROPOSTOS PARA O AJUSTE DE DADOS DE ESPORTES COLETIVOS

Belo Horizonte, 17 de julho de 2022.

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

MARIANA DE CASTRO PASQUALINI

COMPARAÇÃO DO DESEMPENHO DE MODELOS PROPOSTOS PARA O AJUSTE DE DADOS DE ESPORTES COLETIVOS

Trabalho apresentado à banca examinadora da UFMG, como requisito para a obtenção do título de bacharel em Estatística sob a orientação do professor Cristiano de Carvalho Santos.

Belo Horizonte, 17 de julho de 2022.

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

Resumo

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 e Stevenson (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 e Ntzoufras (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 e Blangiardo (2010). Neste trabalho, são implementados, ajustados e comparados modelos baseados na distribuição Poisson para dados do Campeonato Brasileiro de 2019, obtido no repositório cartola no Github utilizando o software Stan e Rstan para inferência bayesiana. O desempenho dos modelos para dados reais são verificados por meio de uma comparação gráfica dos pontos acumulados ao longo da temporada, também é calculado o erro quadrático médio entre a pontuação estimada pelo modelo e observada na competição. Além disso, é utilizado o critério de informação LOO para seleção de modelos.

Sumário

Lista de Figuras

Lista de Tabelas

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 e Stevenson (1988) apresentam as diferenças entre a Binomial Negativa e Poisson para modelar o placar de partidas de futebol. Mais recentemente, podemos citar Diniz et al. (2019) com a proposta de um modelo multinomial e Tsokos et al. (2019) com o modelo Bradley-Terry, no contexto de aprendizado de máquina.

Tais modelos desconsideram uma estrutura de correlação entre os gols de cada oponente. Karlis e Ntzoufras (2003) sugerems 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 e Blangiardo (2010). Neste trabalho, são implementados, ajustados e comparados modelos baseados na distribuição Poisson para dados do Campeonato Brasileiro de 2019, obtidos no repositório caRtola no Github utilizando o software Stan e RStan para inferência bayesiana.

2. Metodologia

Para o problema de contagem do número de gols em uma partida, o modelo mais comum é baseado na distribuição de Poisson, discreta, representando o número de eventos ocorridos em um intervalo de tempo. Esta distribuição 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. Todos os modelos que serão apresentados aqui são baseados nessa distribuição e ajustados aos dados do Campeonato Brasileiro ou "Brasileirão" do ano de 2019. Apesar disso, há outros modelos baseados em diferentes distribuições de probabilidade.

2.1 Modelo Poisson misto

Baio e Blangiardo (2010) sugerem um modelo hierárquico de efeitos aleatórios para os gols marcados em uma determinada partida. No modelo proposto, o número de gols realizados por cada equipe segue uma distribuição Poisson condicionalmente independentes, em que a correlação é incluída por meio dos efeitos aleatórios.

Visto que o vetor $\mathbf{y} = (y_{g1}, y_{g2})$ como um vetor de contagens, podemos assumir

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

isto é, 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 ($j = 1$) e o visitante ($j = 2$).

Utilizando a função de ligação log e um modelo 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.

2.1.1 Distribuições a priori

Considerando que o modelo proposto segue a abordagem bayesiana, os efeitos aleatórios são tratados da mesma forma que parâmetros de interesse e é apropriado definir uma distribuição à priori para cada um deles. As distribuições a priori sugeridas por Baio e Blangiardo (2010) são:

$$home \sim Normal(0, 10),$$
$$att_t \sim Normal(\mu_{att}, \sigma_{att}),$$
$$def_t \sim Normal(\mu_{def}, \sigma_{def}).$$

Sendo t o índice de cada um dos times do campeonato.

Além disso, é necessário definir também as distribuições a priori dos hiperparâmetros. Para as médias, são definidas prioris pouco informativas $\mu_{att} \sim Normal(0, 10)$ e $\mu_{def} \sim Normal(0, 10)$.

Conforme demonstrado por Gelman et al. (2008) a priori não-informativa recomendada para o desvio padrão é uma Cauchy, portanto assumimos $\sigma_{att} \sim Cauchy(0, 2.5)$ e $\sigma_{def} \sim Cauchy(0, 2.5)$.

Para garantir a identificabilidade do modelo, os autores Blaio e Blangiardo (2010) 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 em comparação ao time de referência. Essa alternativa foi implementada neste trabalho e a restrição é dada por:

$$att_T = 0,$$
$$def_T = 0.$$

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

2.2 Modelo Poisson misto com mistura

Segundo os autores Baio e Blangiardo (2010), o primeiro modelo tem a tendência de subestimar a pontuação dos times bons e superestimar os times ruins. A partir disso, é proposto um modelo de mistura com 3 componentes, representando categorias das habilidades do time.

O ataque e defesa seguem uma distribuição t-Student 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 e Blangiardo (2010)).

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

2.2.1 Distribuições a priori

Novamente, define-se distribuições a priori para os parâmetros e hiperparâmetros do modelo. A probabilidade π do ataque ou defesa de um time pertencer ao grupo 1, 2 ou 3 é igual para todos os times, então $\pi_{att} \sim Dirichlet(1,1,1)$ e $\pi_{def} \sim Dirichlet(1,1,1)$.

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 pertence. Para o primeiro grupo, $\mu_1^{att} \sim Normal(0,10)[-3,0]$ e $\mu_1^{def} \sim Normal(0,10)[0,3]$, o segundo *cluster* tem $\mu_2^{att} \sim Normal(0, 0.01)$ e $\mu_2^{def} \sim Normal(0, 0.01)$ e, por fim, para o terceiro grupo definimos as distribuições $\mu_3^{att} \sim Normal(0,10)[0,3]$ e $\mu_3^{def} \sim Normal(0,10)[-3,0]$.

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. Como é uma extensão do modelo 1, as demais especificações do modelo permanecem iguais ao modelo inicial.

2.3 Modelo misto Poisson bivariado

2.3.1 Poisson bivariada

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, partindo do princípio que o número de gols de um time não afeta o número de gols do adversário. 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 e 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 $\mathbf{BP}(\lambda_1, \lambda_2, \lambda_3)$. A partir disso, 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 λ_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, a quantidade marcada pelo oponente 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.

2.3.2 Modelos de regressão

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

$$Y_{i1} \sim Poisson(\lambda_{1i}),$$
$$Y_{i2} \sim Poisson(\lambda_{2i})$$

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

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

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

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

O qual γ_j é uma variável *dummy*, indicando quais parâmetros serão incluídos no modelo de interesse. Além disso, diferentemente dos modelos anteriores, neste temos a inclusão de um intercepto μ.

No artigo original, ataque e defesa são tratados como efeitos fixos e estimados via algoritmo EM, 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. Neste trabalho, os efeitos foram tratados como aleatórios e definiu-se distribuições a priori para cada um deles.

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

$$att_T = 0,$$
$$def_T = 0.$$

2.3.3 Distribuições a priori

A escolha das distribuições a priori dos modelos derivados da Poisson Bivariada segue o mesmo princípio do modelo 1: prioris pouco informativas.

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

Baseado nas definições apresentadas acima, é possível definir modelos distintos, cada um incluindo diferentes efeitos na medida de correlação entre as variáveis aleatórias que representam o número de gols do mandante e visitante. Isso é feito pelas combinações dos valores de γ_1 e γ_2.

3. Estudos de simulação

Estudos com dados simulados foram realizadas para os modelos Poisson misto e Poisson bivariado com $\gamma_1 = \gamma_2 = 0$, com o objetivo de se checar a estimação correta dos parâmetros pelos modelos implementados. Ambas foram feitas com apenas um cenário.

3.1 Modelo Poisson misto

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. Os parâmetros do modelo usados para geração dos bancos de dados foram $home = 0.13$, $\mu_{att} = 0.05$, $\mu_{def} = 0.08$, $\sigma_{att} = 0.56$ e $\sigma_{def} = 0.52$.

As simulações foram realizadas com apenas 01 cadeia e 5000 interações, sendo 2500 de aquecimento. O gráfico traceplot, apresentado na Figura 1, mostra que a cadeia converge e consegue caminhar pelo espaço paramétrico.

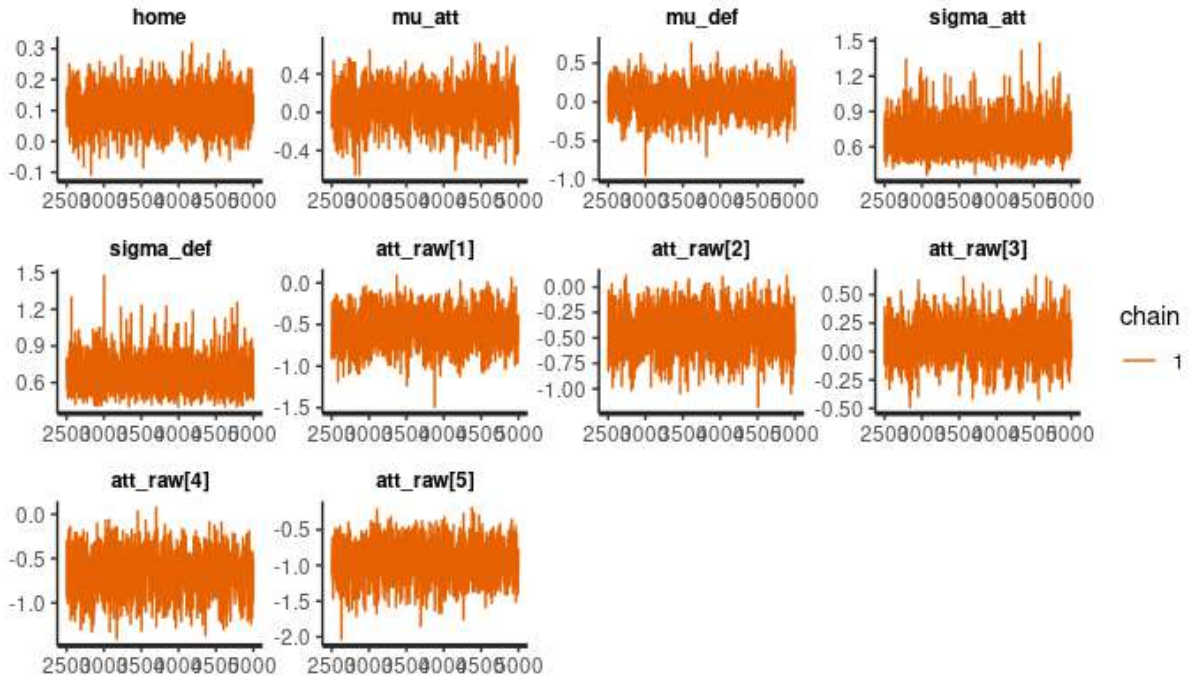

Figura 1: Traceplot para análise de convergência - Modelo Poisson misto 1

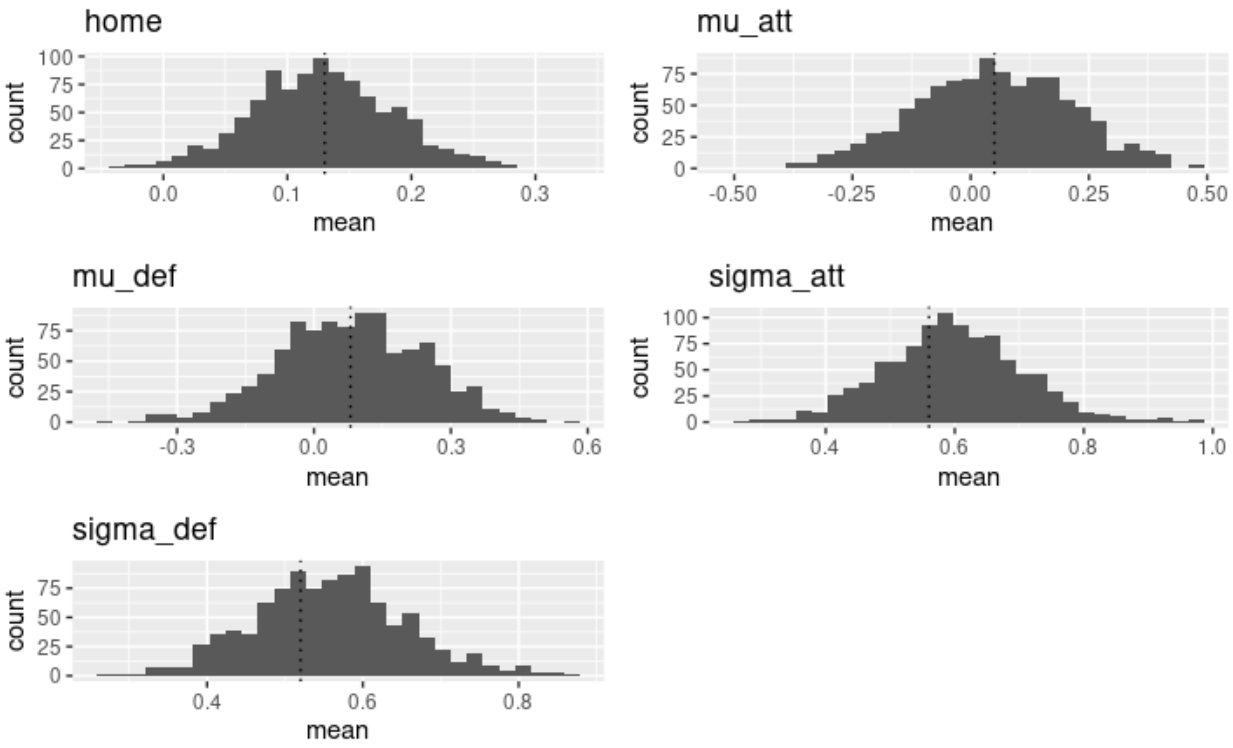

Figura 2: Histograma das estimativas para cada parâmetro ao utilizarmos a média a posteriori - Modelo Poisson misto

A partir da Figura 2, são apresentados histogramas das estimativas das médias a posteriori para alguns parâmetros. Nessa imagem, observa-se que as distribuições da média da distribuição a posteriori dos parâmetros estão centradas em torno dos valores reais. A simulação durou aproximadamente uma hora para terminar a execução.

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 $\hat{R} = 1.002963$.

3.2 Poisson bivariado com $\gamma_1 = \gamma_2 = 0$

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

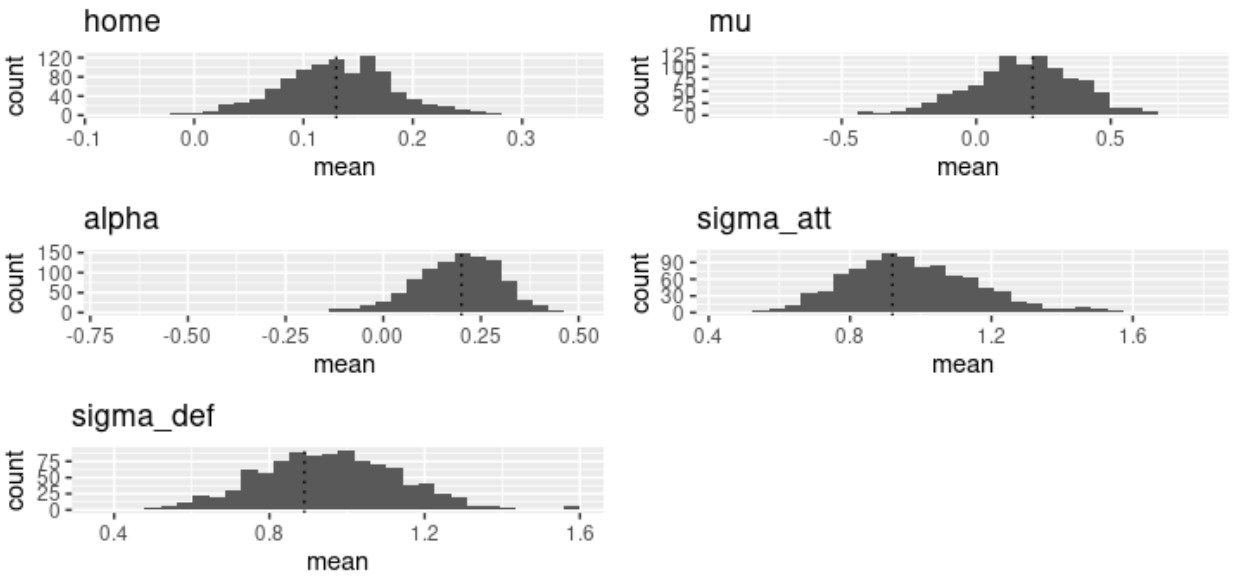

Figura 3: Histograma das estimativas para cada parâmetro ao utilizarmos a média a posteriori - Poisson bivariado

Na Figura 3, a partir dos resultados dos histogramas obtidos na simulação, tem-se que o modelo estima corretamente os parâmetros. Para essa execução, foram aproximadamente sete horas. 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 $\hat{R} = 1.002963$.

4. Ajuste para dados reais

Para verificar o comportamentos do modelos para dados reais, foi obtido o conjunto de dados do Campeonato Brasileiro do ano de 2019. Nos artigos originais, os modelos são ajustados para dados da Série A do Campeonato italiano. Foi escolhido 2019 por se tratar da competição mais recente antes da pandemia.

O dados foram disponibilizados por Gomide e Gualberto no repositório caRtola, disponível no Github, com o seguinte formato:

Tabela 1: Formato do conjunto de dados do Campeonato Brasileiro 2019

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

Conforme apresentado na Tabela 1, as colunas *home_team_index* e *away_team_index* foram criadas atribuindo um valor inteiro ordinal para cada time, seguindo a notação do modelo.

O que foi modelado até agora é o número de gols marcados pelas equipes em uma determinada partida. Porém, em um campeonato, também há o interesse na **pontuação** de cada time ao longo das rodadas.

A partir da distribuição preditiva a posteriori, é possível prever o placar da partida e determinar:

- 3 pontos para o time vitorioso
- 1 ponto para casos de empate
- 0 pontos para derrota

Além da interpretação dos efeitos estimados, a pontuação acumulada será comparada entre o que foi observado no campeonato e o estimado pelos seis modelos.

4.1 Ajuste dos modelos

Nesta seção, apresentamos os resultados obtidos após ajustar os modelos apresentados na seção 2. Para simplificação, denotamos o modelo Poisson misto como modelo 1, o modelo Poisson misto com mistura como modelo 2, apresentados nas seções 2.1 e 2.2, respectivamente. O modelo 3 refere-se ao modelo misto Poisson bivariado com $\gamma_1 = \gamma_2 = 0$ e o modelo 4 com $\gamma_1 = 1, \gamma_2 = 0$. Por fim, os modelo 5 e 6, respectivamente, representam, os modelos mistos Poisson bivariado com $\gamma_1 = 1 = \gamma_2 = 1$ e $\gamma_1 = 0, \gamma_2 = 1$.

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

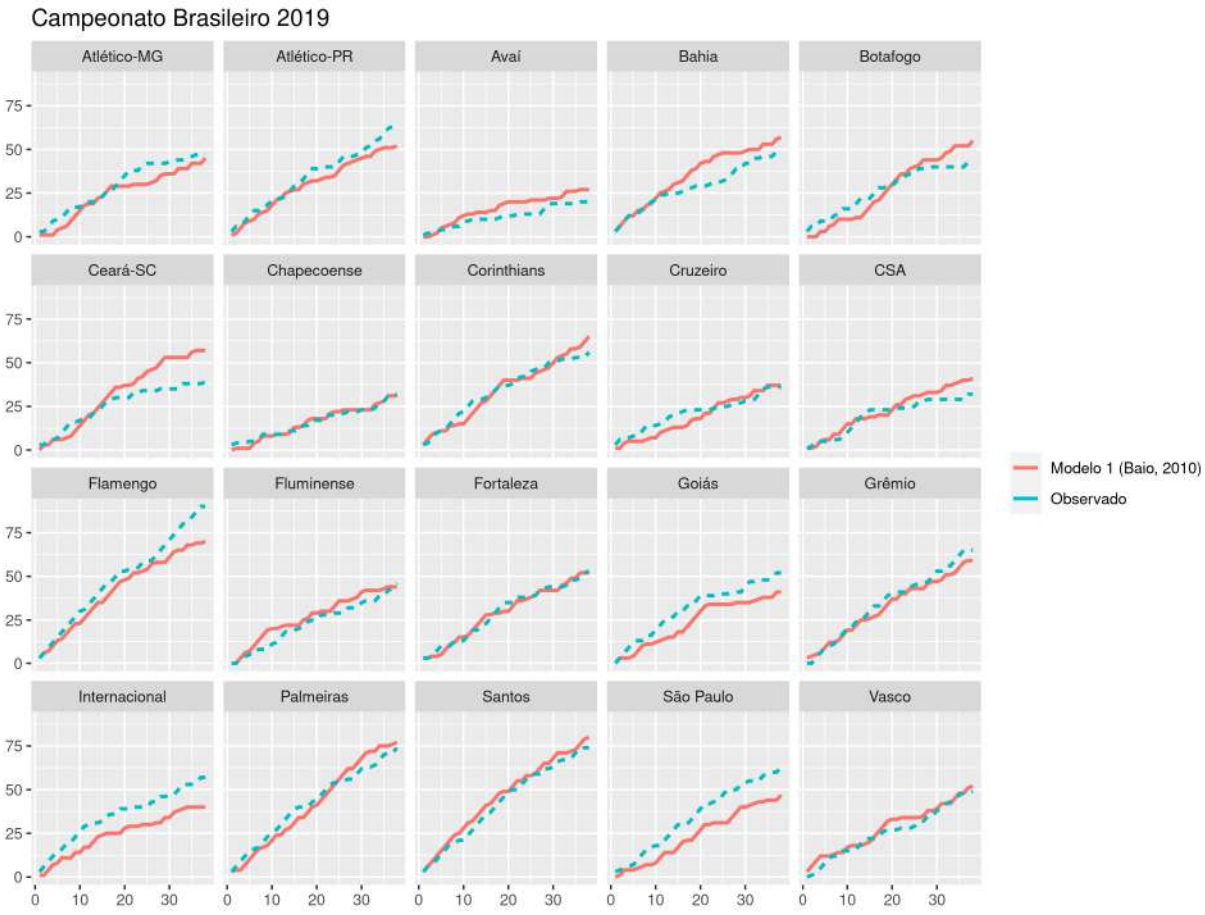

Figura 4: Comparação da pontuação acumulada observada e prevista, segundo o modelo 1

De acordo com a Figura 4, Santos foi time com maior pontuação atribuída pelo modelo (80 pontos), estimando bem próximo da pontuação obtida (74 pontos) e, assim, sendo o campeão segundo o modelo. O modelo errou

apenas 1 dos 4 melhores times, colocando o Corinthians no lugar do grêmio. Para os rebaixados, o modelo incluiu o Internacional e não rebaixaria o CSA, segundo as estimativas de pontuação.

Na Tabela 2, apresentamos os efeitos de ataque e defesa são interpretados com base na referência, definida anteriormente como o último time pelo seu índice, que neste caso é o Fortaleza.

Tabela 2: Efeitos de ataque estimados através do 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

Segundo a Tabela 2, vemos que o 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. Considerando o Flamengo e o Fortaleza jogando o mesmo jogo como mandante e contra o mesmo adversário, a ocorrência média de gols do Flamengo é $\exp(0.546) = 1.73$ vezes a do Fortaleza. Já o Cruzeiro apresenta um efeito de ataque negativo e seu intervalo de credibilidade de 90% é [-0.767; -0.124], com a ocorrência média de gols do time de $\exp(-0.438) = 0.645$ vezes a do Fortaleza. Em relação a linha de base, observa-se que a ocorrência média de gols para o Flamengo é maior que a do Cruzeiro, coerente com o resultado real do Brasileirão.

Tabela 3: Efeitos de defesa estimados através do 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

Analisando o efeito de defesa a partir da Tabela 3, os times com maior média apresentam maior propensão de conceder gols, lembrando do sinal positivo no preditor linear do modelo 1 $\log \theta_{g2} = att_{a(g)} + def_{h(g)}$. Por isso, os piores times no campeonato apresentam sinais positivos e os melhores, sinais negativos. A equipe do São Paulo tem uma propensão de conceder gols ao adversário de $\exp(-0.234) = 0.791$ vezes a do Fortaleza, enquanto o Avaí tem $\exp(0.160) = 1.173$ vezes em relação a mesma linha de base.

Tabela 4: Efeitos de casa estimado através do modelo 1

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

É possível observar o efeito de jogar em casa na Tabela 4. Conforme o esperado, o efeito de jogar em casa é positivo, contribuindo para a ocorrência média de gols pelo time da casa, com intervalo de credibilidade de 90% entre 0.299 e 0.53.

Já para o modelo 2, antes da análise dos resultados o modelo é necessário apresentar a convergência das cadeias para os dados reais. Diferentemente das cadeias dos outros modelos ajustados, que podem ser encontradas no Apêndice, para este modelo em específico houve problemas de convergência. Foram usadas duas cadeias para estimação das distribuições a posteriori, o parâmetro *thin* igual a 5, descartando tal valor de amostras, e 10000 iterações.

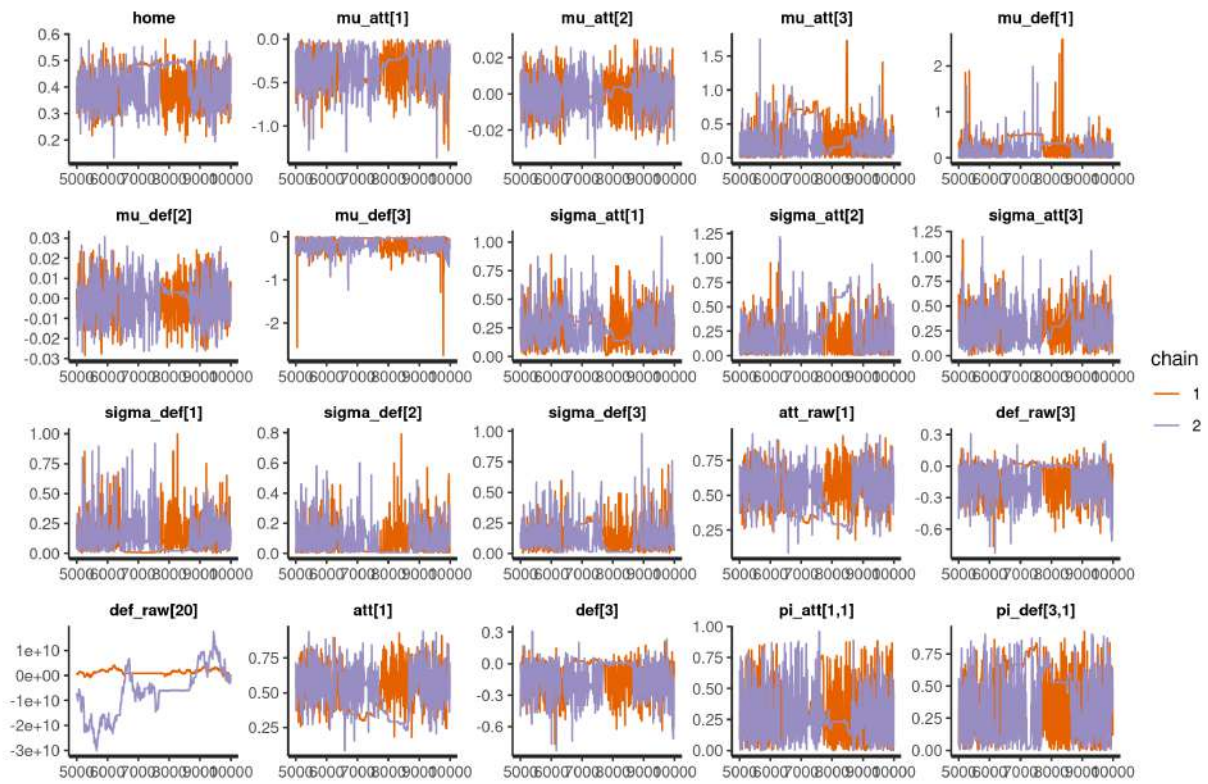

Figura 5: Gráfico de traço da cadeia a posteriori de alguns parâmetros ao ajustar o modelo 2

De acordo com os *traceplots* de alguns parâmetros apresentados na Figura 5, observamos 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 no trabalho, mas a inferência pode estar comprometida.

Sendo assim, foi feito o mesmo gráfico de pontuação acumulada ao longo do campeonato. Para cada time, temos o gráfico apresentado na Figura 6:

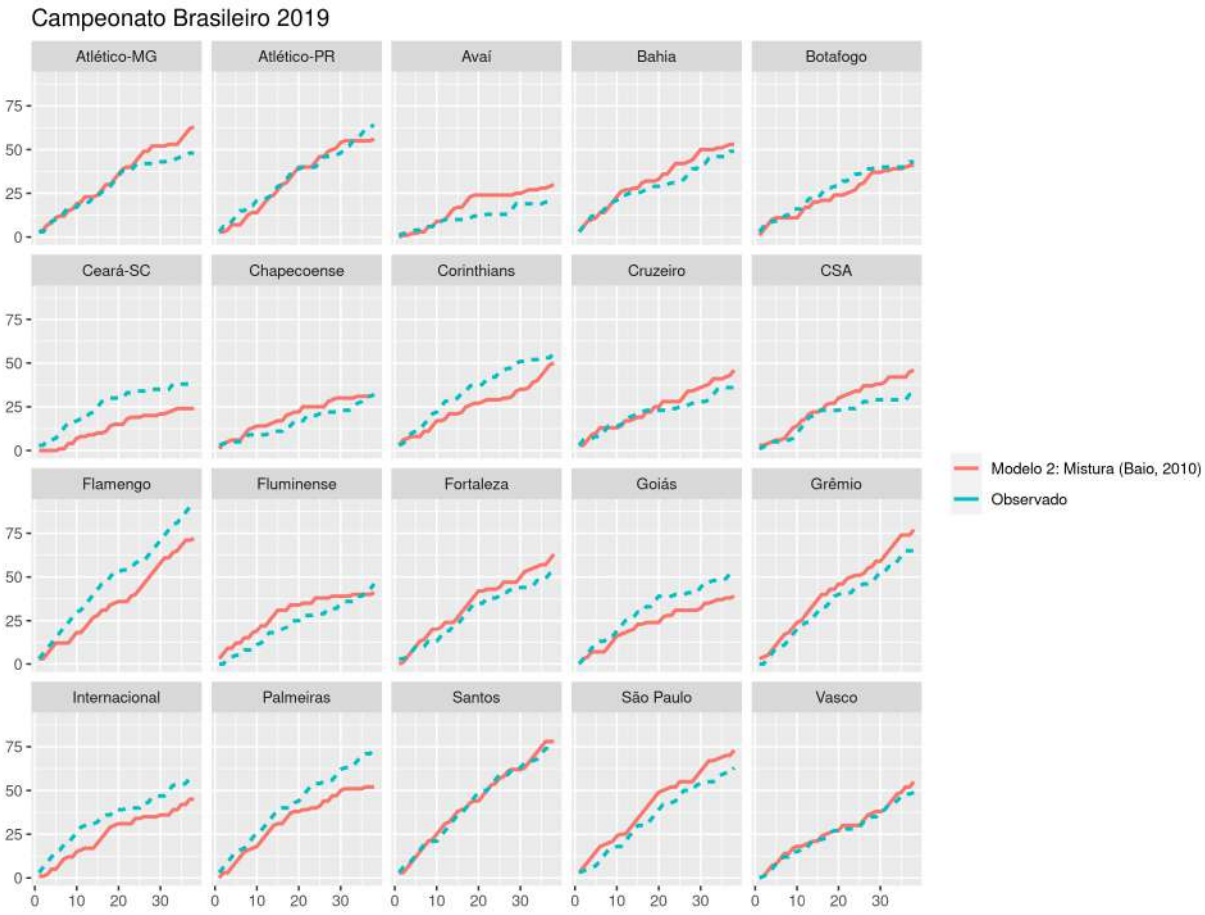

Figura 6: Comparação da pontuação acumulada observada e prevista, segundo o modelo 2

Ao analisarmos os resultados obtidos no ajuste do segundo modelo, é necessário ter cautela, pois é importante considerar uma possível má especificação do modelo, por isso a análise deve ser realizada com cautela. Em Baio (2010) há uma divergência entre a definição do modelo no texto e o código, além de não ter o diagnóstico dos métodos MCMC. Com essa consideração, o modelo consegue estimar bem o *top* 4 do campeonato, acertando todos os times. Já para os rebaixados, acerta apenas metade dos clubes. Os efeitos estimados do modelo estão no Apêndice.

Na Figura 7, encontra-se o gráfico da pontuação acumulada para o modelo 3:

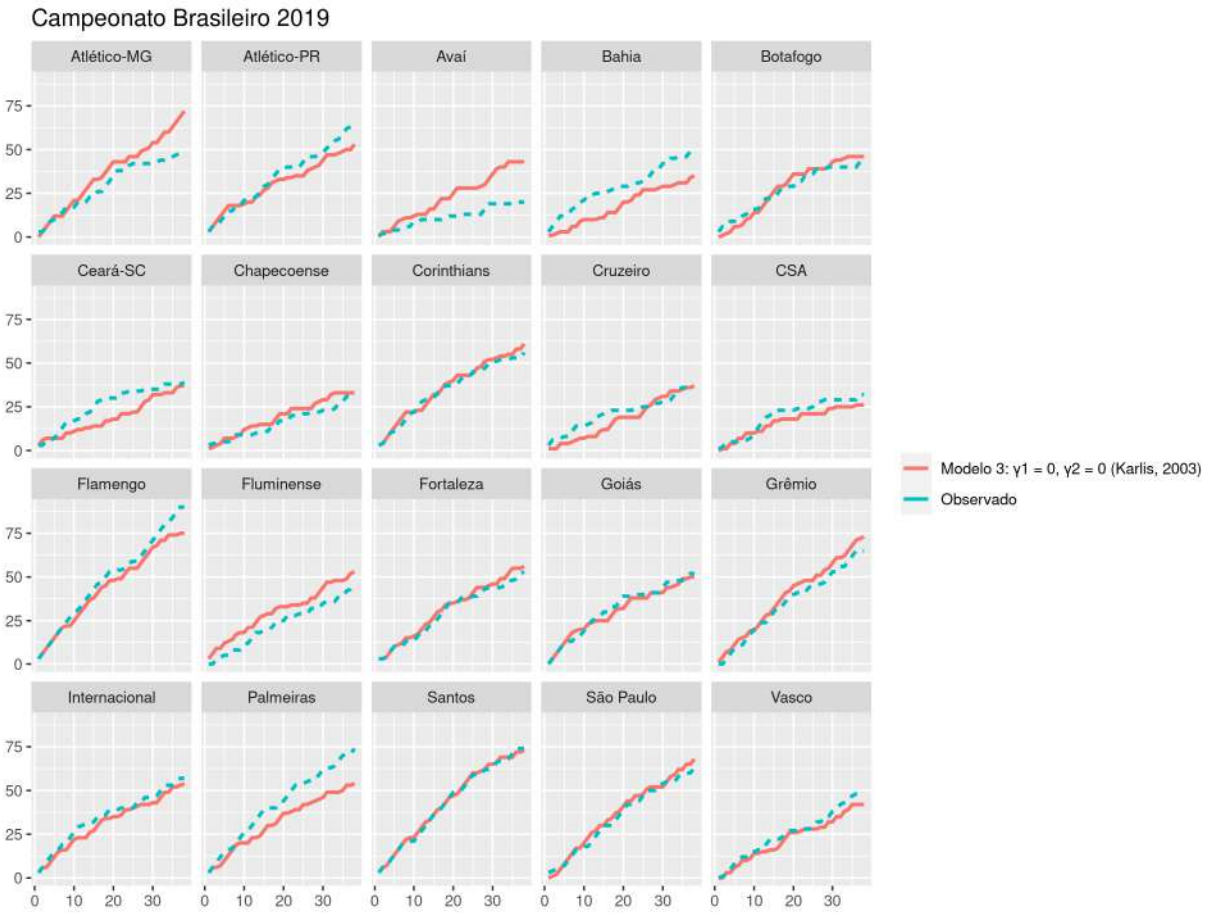

Figura 7: Comparação da pontuação acumulada observada e prevista, segundo o modelo 3

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. Apesar disso, o modelo atribui as menores pontuações aos times rebaixados e deixa de fora *top* 4 apenas o Palmeiras, colocando o Atlético-MG no topo da tabela.

Tabela 5: Efeito de ataque de cada time, estimado pelo Modelo 3

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	1.205	1.162	0.355	0.691	1.846
Botafogo	-0.542	-0.470	0.540	-1.516	0.142
Corinthians	-0.038	-0.013	0.425	-0.740	0.617
Bahia	-0.025	0.012	0.416	-0.704	0.583
Fluminense	-0.553	-0.444	0.569	-1.650	0.167
Vasco	-0.355	-0.272	0.532	-1.215	0.289
Palmeiras	0.689	0.660	0.341	0.186	1.266
São Paulo	-0.245	-0.192	0.449	-1.071	0.374
Santos	0.731	0.696	0.352	0.226	1.356
Atlético-MG	0.117	0.131	0.391	-0.497	0.722
Cruzeiro	-0.932	-0.846	0.593	-2.043	-0.145
Grêmio	0.729	0.711	0.333	0.228	1.311
Internacional	0.083	0.092	0.390	-0.539	0.718
Goiás	0.180	0.178	0.368	-0.404	0.781
Atlético-PR	0.214	0.228	0.390	-0.358	0.814
Avaí	-1.278	-1.179	0.657	-2.489	-0.416
Chapecoense	-0.627	-0.549	0.529	-1.621	0.098
CSA	-0.875	-0.789	0.564	-1.931	-0.113
Ceará-SC	-0.335	-0.278	0.497	-1.194	0.339
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 6: Efeito de defesa de cada time, estimado pelo Modelo 3

	Time	Média	Mediana	Desvio-padrão	5%	95%
21	Flamengo	-0.171	-0.138	0.290	-0.686	0.248
22	Botafogo	0.043	0.041	0.238	-0.348	0.427
23	Corinthians	-0.275	-0.238	0.284	-0.794	0.122
24	Bahia	-0.087	-0.072	0.256	-0.525	0.303
25	Fluminense	0.059	0.059	0.239	-0.339	0.450
26	Vasco	0.067	0.066	0.236	-0.319	0.458
27	Palmeiras	-0.241	-0.220	0.266	-0.711	0.142
28	São Paulo	-0.360	-0.319	0.301	-0.892	0.068
29	Santos	-0.299	-0.266	0.312	-0.869	0.122
30	Atlético-MG	0.051	0.053	0.258	-0.390	0.461
31	Cruzeiro	0.055	0.055	0.240	-0.326	0.440
32	Grêmio	-0.190	-0.160	0.284	-0.718	0.212
33	Internacional	-0.166	-0.140	0.279	-0.677	0.241
34	Goiás	0.547	0.528	0.265	0.140	1.008
35	Atlético-PR	-0.241	-0.217	0.280	-0.707	0.147
36	Avaí	0.418	0.404	0.246	0.040	0.848
37	Chapecoense	0.098	0.100	0.251	-0.313	0.505
38	CSA	0.244	0.239	0.246	-0.122	0.651
39	Ceará-SC	-0.030	-0.020	0.248	-0.460	0.355
40	Fortaleza	0.000	0.000	0.000	0.000	0.000

Os coeficientes podem ser interpretados da mesma maneira, tomando o efeito de ataque do Santos na Tabela 4, a ocorrência média de gols é de $\exp(0.731) = 2.077$ vezes a do Fortaleza. Para o efeito de defesa, apresentado na Tabela 5, a propensão de conceder gols ao adversário do Goiás é de $\exp(0.547) = 1.728$ vezes a do Fortaleza.

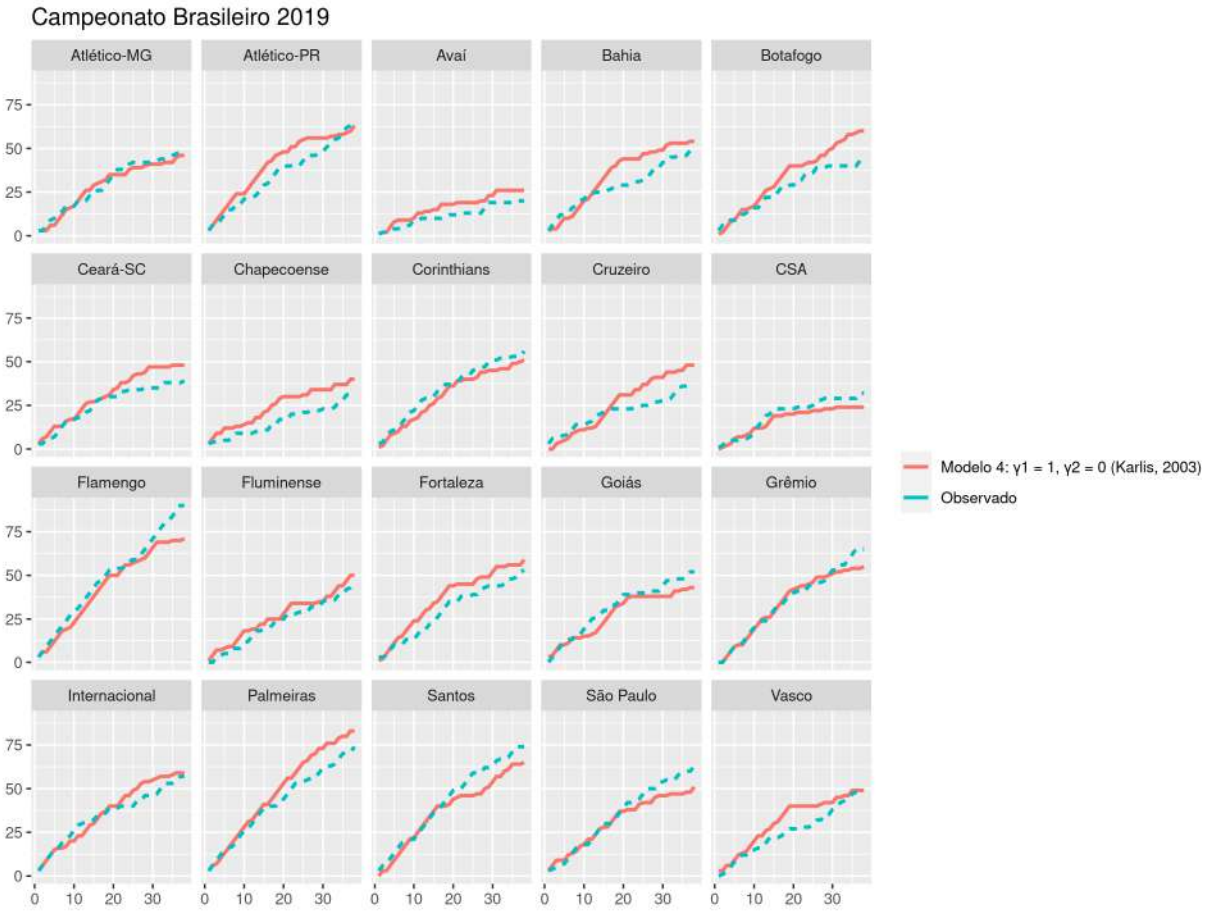

Figura 8: Comparação da pontuação acumulada observada e prevista, segundo o modelo 4

Observando a Tabela 8, vemos que, diferentemente dos modelos anteriores, o quarto modelo consegue acertar o time carioca como campeão, com 79 pontos. Apesar da estimativa afastar em algumas rodadas, a pontuação final se aproxima muito do observado. Os times com as piores estimativas são o Ceará e Avaí.

Apresentamos o gráfico da pontuação acumulada também para o modelo 5.

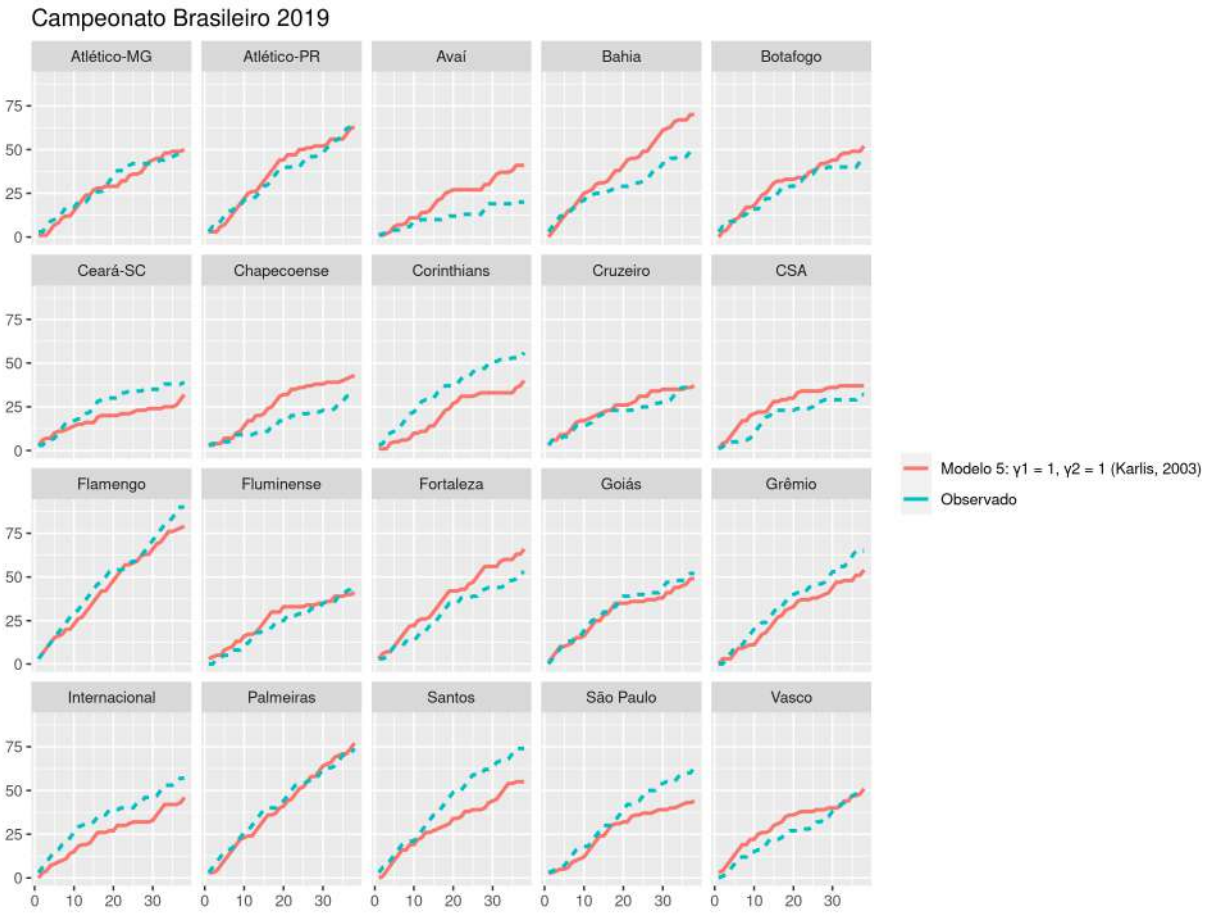

Figura 9: Comparação da pontuação acumulada observada e prevista, segundo o modelo 5

Pelo gráfico da Figura 9, há um indicativo desse modelo ser o pior até o momento. As curvas se afastam notavelmente para vários times, como Santos, Avaí e São Paulo. Dos quatro melhores times, o modelo acertou apenas Flamengo e Palmeiras. Nos piores, também acertou apenas o Cruzeiro e CSA, estimando erroneamente que Corinthians e Ceará também seriam rebaixados.

Para o sexto modelo, o gráfico da pontuação acumulada e prevista está na Figura abaixo:

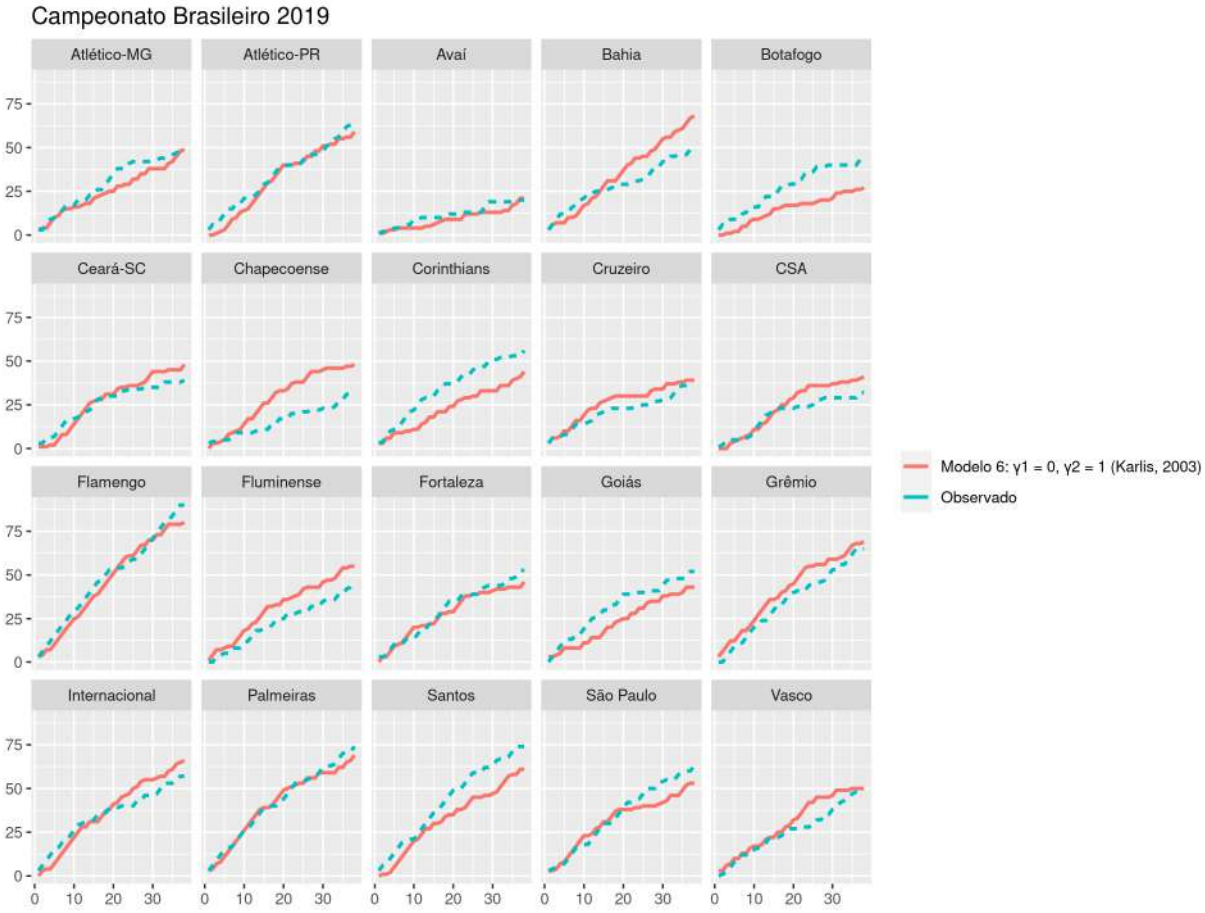

Figura 10: Comparação da pontuação acumulada observada e prevista, segundo o modelo 6

Analisando a Figura 10, o último modelo também segue um comportamento parecido com os anteriores, estimando muito bem a pontuação do Palmeiras e encontrando com a pontuação final do Cruzeiro e o Atlético Mineiro. As estimativas mais distantes ficam para o Botafogo, Corinthians e Chapecoense. Esse modelo erra apenas um time daqueles com pior desempenho, incluindo Botafogo e excluindo a Chapecoense dos rebaixados. Já os melhores, acerta 3 dos times, estimando que a pontuação do Bahia seria uma das mais altas e removendo o Santos do *top* 4.

4.2 Comparação dos modelos

Com o objetivo de escolher o melhor modelo, foram definidas duas formas de comparação: erro quadrático médio da pontuação no campeonato e a estatística LOO, obtida por meio de um método de validação cruzada.

Inicialmente, pode-se fazer uma comparação visual das pontuações acumuladas no campeonato de todos os modelos ajustados e o observado, para compreender o comportamento geral das curvas. Isso é mostrado na Figura 11.

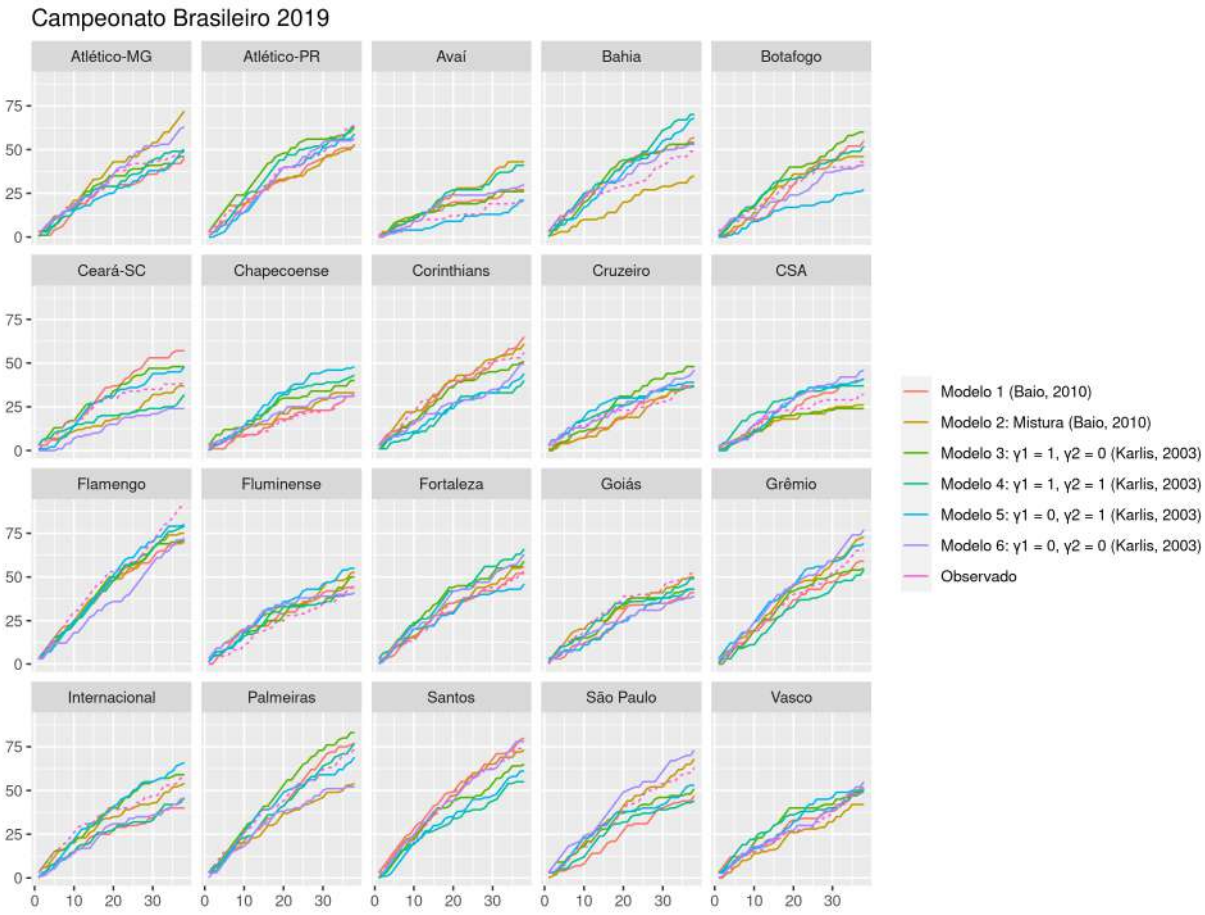

Figura 11: Comparação da pontuação acumulada observada e prevista, de acordo com todos os modelos definidos

Na Figura 11, o tracejado representa a pontuação observada. Observa-se que alguns modelos subestimam o desempenho de alguns times e outros são superestimados. Para auxiliar a visualização, é fundamental resumir a informação do desempenho dos modelos.

4.2.1 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. Calculando o erro quadrático médio, obtém-se o quanto o modelo se distancia, em média, do valor real.

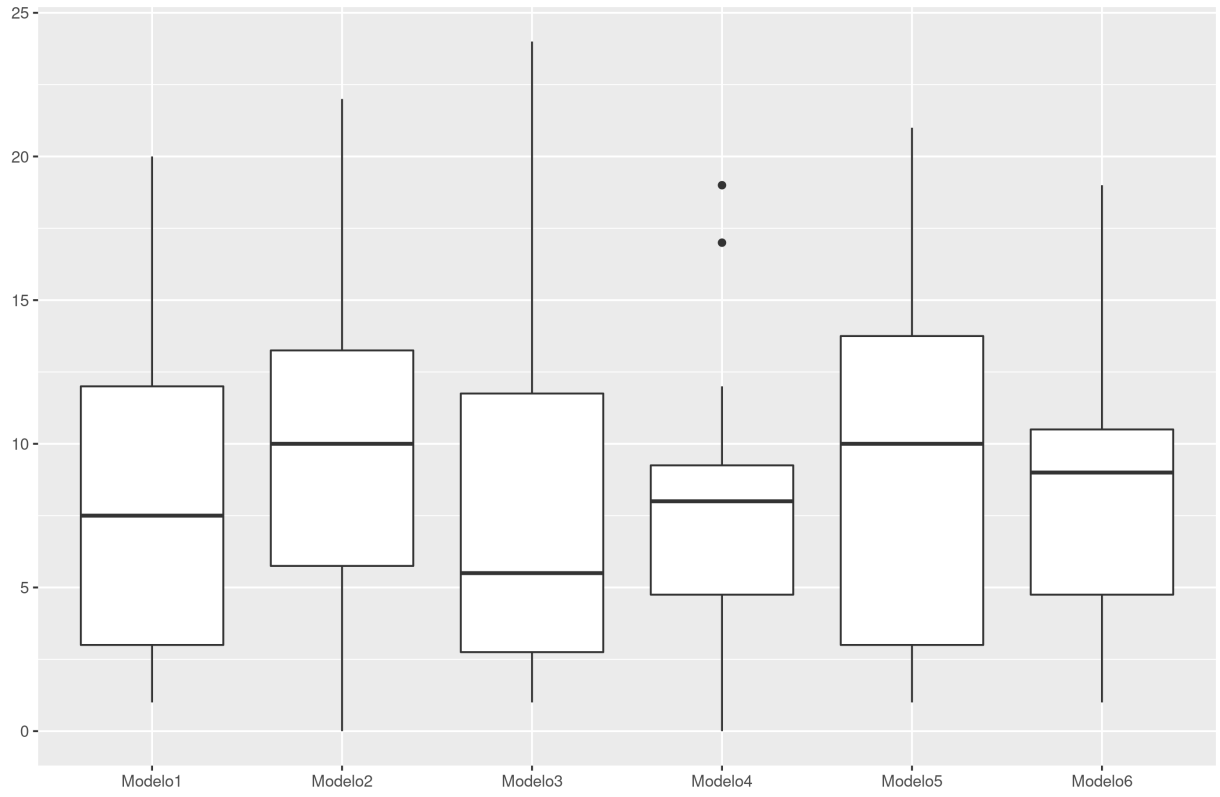

Figura 12: Boxplot comparando o erro quadrático médio da pontuação estimada pelos modelos

No boxplot da Figura 12, temos o erro quadrático médio da pontuação de cada time estimada pelos seis modelos. As medianas mais próximas de 0 identificam o melhor modelo. A partir disso, nesse critério, o melhor modelo seria o terceiro, mas as caixas apresentam uma grande amplitude que representa uma dispersão entre o estimado e o valor real da pontuação. Sugere-se então uma medida mais robusta para identificação do melhor modelo.

4.2.2 Critério de informação LOO e validação cruzada

Vehtari et al. (2015) propõe o cálculo de um critério de informação LOO-CV (*leave-one-out cross-validation*), que é feito baseado na esperança da densidadade preditiva do modelo. É definido

$$elpd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i})$$

onde $p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i}d\theta)$ é a densidade preditiva do modelo, considerando todas as observações exceto a que se quer fazer a predição. Daí, é possível calcular o critério de informação $LOO_{ic} = -2 \times elpd_{loo}$. A partir disso, quanto menor a medida, melhor o modelo. Assintoticamente, esse critério de informação é igual ao WAIC (*widely applicable information criterion*).

As médias de LOO obtidos para todos os modelos estão apresentados na Tabela a seguir:

Tabela 7: Médias de LOO obtidas para cada modelo

model	elpd_diff	se_diff	looic	se_looic
Modelo 1	0.000	0.000	2013.498	35.843
Modelo 2	-0.249	1.062	2013.997	35.462
Modelo 3	-1122.868	79.951	4259.234	191.284
Modelo 6	-2631.940	145.988	7277.379	323.430
Modelo 4	-2758.863	151.608	7531.224	334.694
Modelo 5	-4135.550	208.729	10284.598	448.458

Na Tabela 7, a coluna looic indica o critério de informação 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, assim em diante. É importante ressaltar novamente que, apesar da inclusão do modelo 2, como suas cadeias de Markov não estão adequadas, a inferência do modelo está comprometida. De acordo com esse critério, o melhor modelo seria o primeiro, seguido do modelo de mistura (número 2). Já o pior modelo é o de número cinco, que é o mais completo baseado na Poisson Bivariada.

5. Considerações finais

Foi observado o ajuste de 6 modelos para os dados do Campeonato Brasileiro 2019. O melhor modelo ajustado, seguindo o critério de informação LOO-CV, é o modelo hierárquico proposto por Baio e Blangiardo (2010). O modelo Poisson bivariado (modelo 5), com $\gamma_1 = \gamma_2 = 1$ é o pior por este critério. Tais resultados não são coerentes com o EQM da pontuação do campeonato, porém o LOO-CV parece mais completo, no sentido de englobar a verossimilhança de cada observação, enquanto o erro quadrático médio leva em conta apenas o ajuste.

Um possível problema com modelos hierárquicos, como o primeiro modelo, é um efeito de encolhimento (*overshrinkage*), no qual subestima-se valores muito altos e superestima os menores. Esse efeito parece bastante razoável de ser observado no contexto esportivo. A partir disso, Baio e Blangiardo (2010) recomenda um modelo de mistura com três componentes para contornar esse efeito. O modelo de mistura é o segundo melhor de acordo com o critério de informação escolhido, porém existe alguns problemas nas suas estimativas. É interessante avaliar se, a partir do modelo com cadeias bem-comportadas, a mistura teria um desempenho melhor que o modelo hierárquico.

Uma das limitações das variáveis aleatórias que seguem uma Poisson é que $E[X] = Var(X) = \lambda$ e, portanto, se há uma superdispersão nos dados, $Var(X) > E[X]$, o modelo pode não ser tão apropriado. Daí, surge a possibilidade de modelagem assumindo a distribuição Binomial Negativa para a variável resposta ou outra distribuição que considere superdispersão nos dados. Modelos para dados com excesso de zero também podem ser avaliados para os dados em questão.

Há ainda a possibilidade de ajuste para outros anos, como 2020 e 2021, para ver o comportamento do modelo diante de novos times. Uma hipótese levantada e analisada por Benz e Lopez (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, 2021) e polo aquático (Karlis e Ntzoufras, 2003).

Referências

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, 37(2), 253–264.

Gabrio, A. (2021). Bayesian hierarchical models for the prediction of volleyball results. Journal of Applied Statistics, 48(2), 301–321.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381–393.

Pollard, R. (1985). 69.9 goal-scoring and the negative binomial distribution. The Mathematical Gazette, 69(447), 45–47.

Baxter, M., & Stevenson, R. (1988). Discriminating between the poisson and negative binomial distributions: An application to goal scoring in association football. Journal of Applied Statistics, 15(3), 347–354.

Gelman, A., Jakulin, A., Pittau, M., & Su, Y. (2008). A default prior distribution for logistic and other regression models. Annals of Applied Statistics, 2(4), 1360–1383.

Benz, L., & Lopez, M. (2021). Estimating the change in soccer's home advantage during the Covid-19 pandemic using bivariate Poisson regression. AStA Advances in Statistical Analysis, 1–28.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and computing, 27(5), 1413–1432.

Stan Development Team. (2022). Stan Modeling Language Users Guide and Reference Manual, Version 2.21.0.

Marco Henrique de Almeida Inácio. Introdução ao Stan como ferramenta de inferência bayesiana.

R Core Team. (2019). R: A Language and Environment for Statistical Computing.

Gomide, A.. (2022). CaRtola: Extração de dados da API do CartolaFC, análise exploratória dos dados e modelos preditivos em R e Python.

McElreath, R. (2020). Statistical rethinking: a Bayesian course with examples in R and Stan. Taylor and Francis, CRC Press.

6. Apêndice

6.1 Traceplot para diagnóstico das cadeias

Abaixo, os gráficos de traço da cadeia a posteriori dos modelos ajustados com os dados do Campeonato Brasileiro. Como o número de parâmetros é muito grande, apenas alguns foram selecionados.

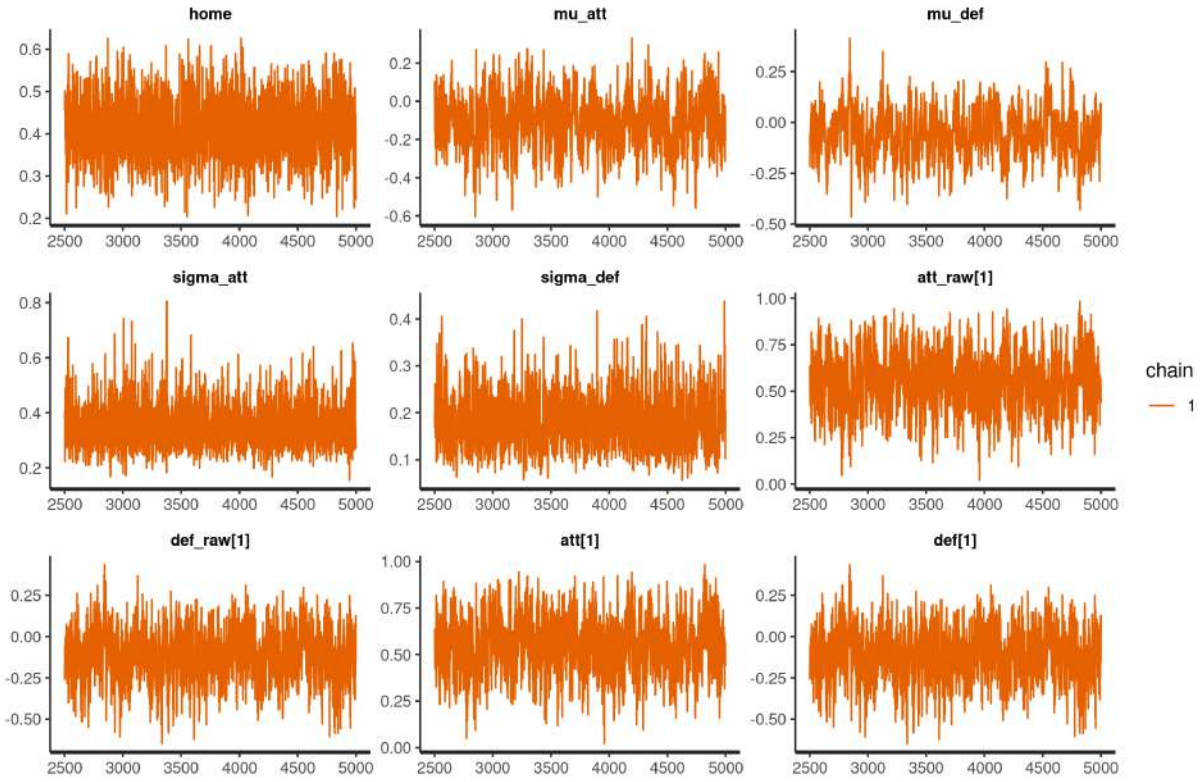

Figura 13: Gráfico de traço da cadeia a posteriori de alguns parâmetros ao ajustar o modelo 1

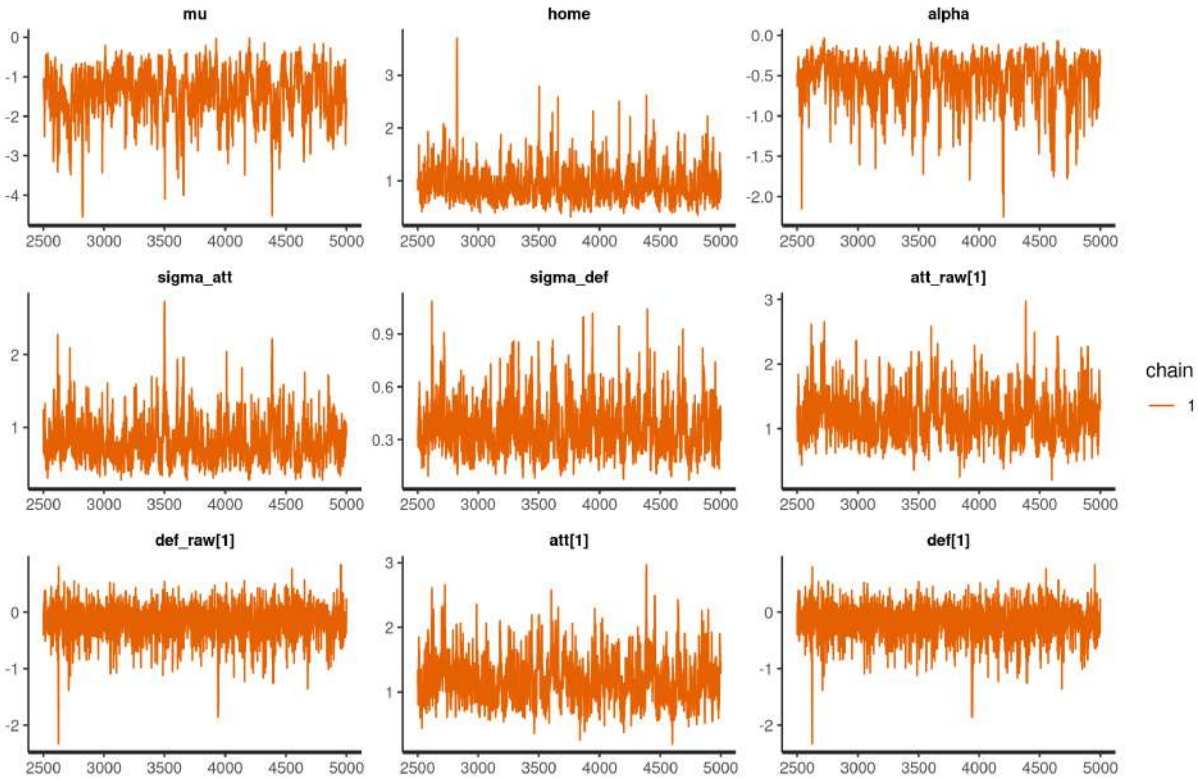

Figura 14: Gráfico de traço da cadeia a posteriori de alguns parâmetros ao ajustar o modelo 2

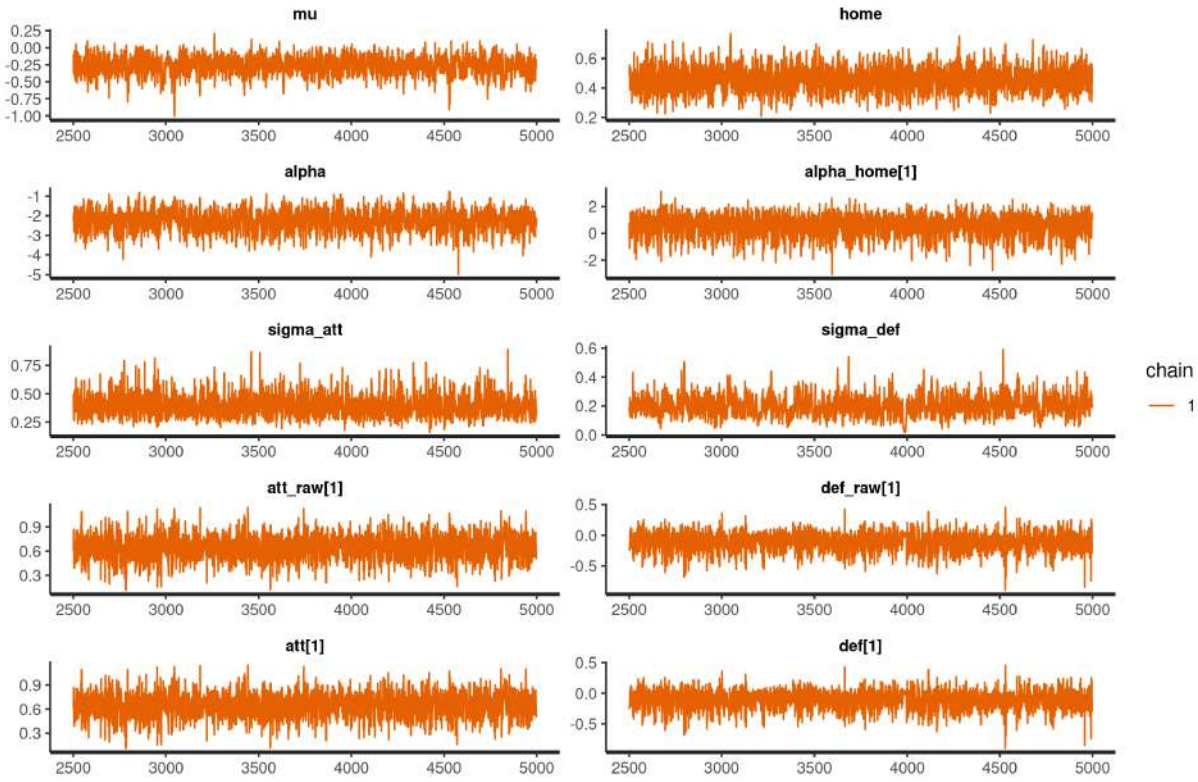

Figura 15: Gráfico de traço da cadeia a posteriori de alguns parâmetros ao ajustar o modelo 3

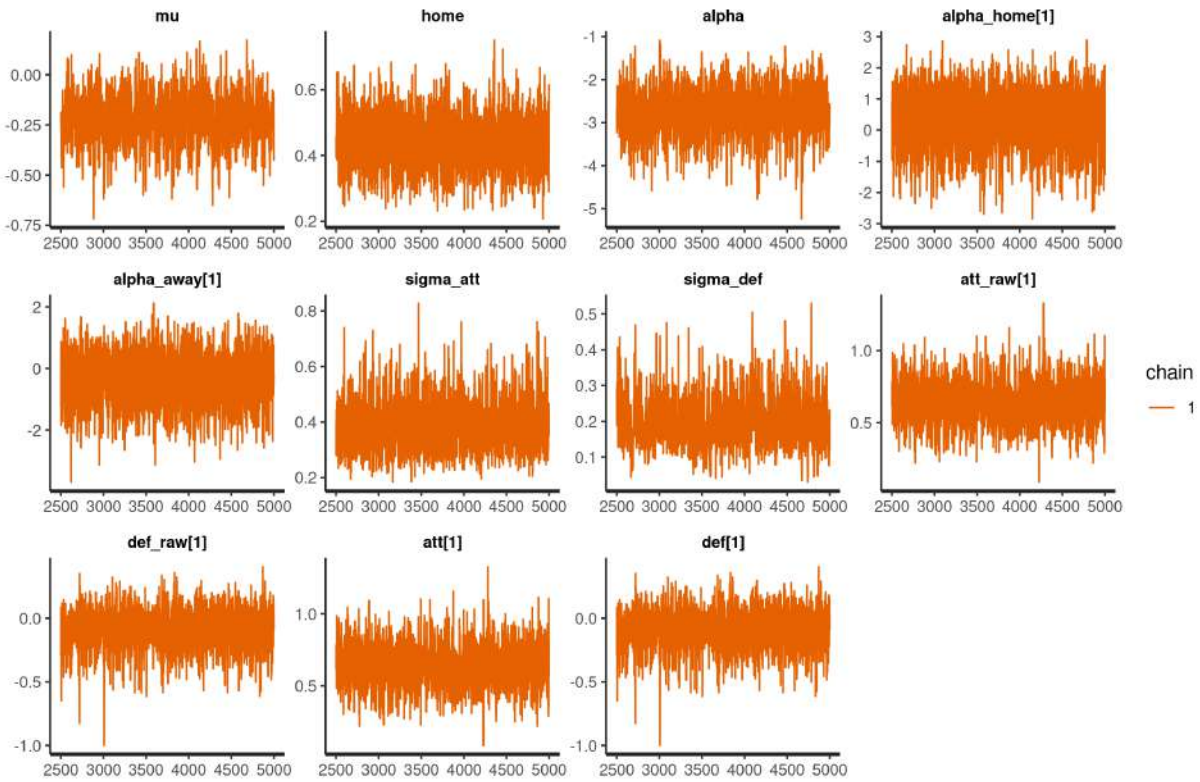

Figura 16: Gráfico de traço da cadeia a posteriori de alguns parâmetros ao ajustar o modelo 4

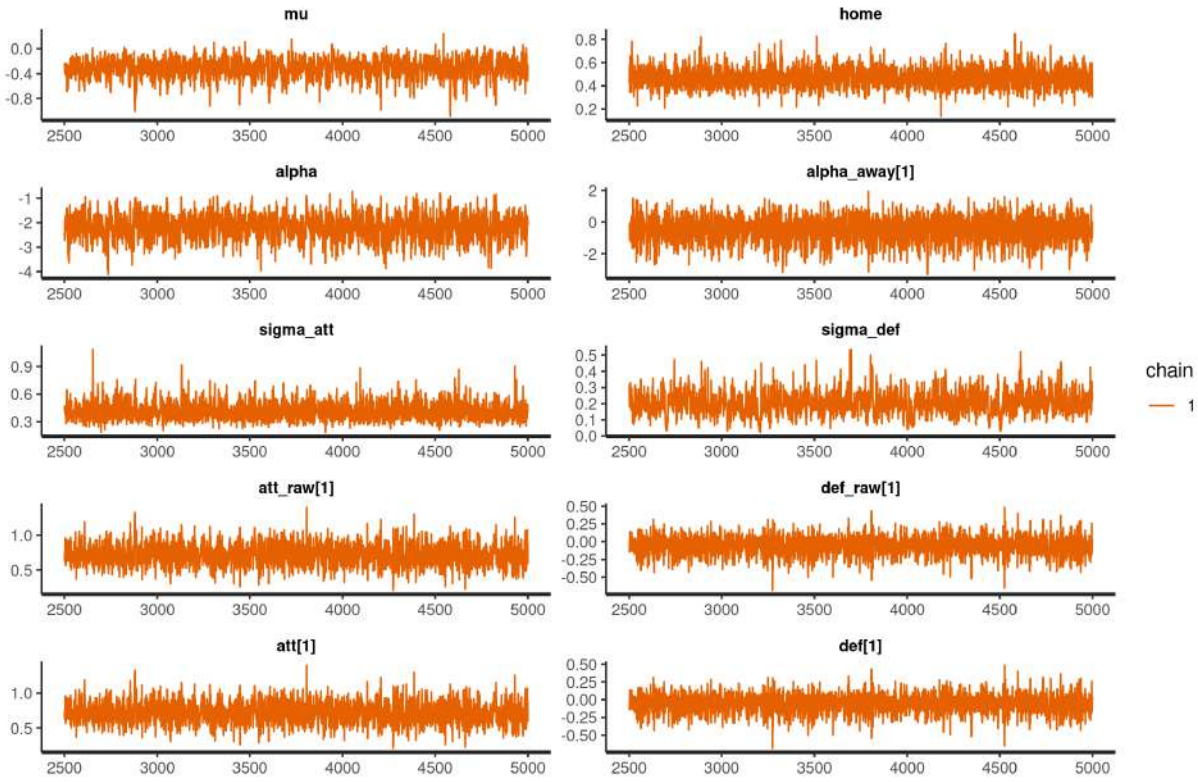

Figura 17: Gráfico de traço da cadeia a posteriori de alguns parâmetros ao ajustar o modelo 5

6.2 Efeitos de ataque e defesa estimados

Nesta seção, são apresentadas as Tabelas dos efeitos de ataque e defesa estimados pelos modelos definidos neste trabalho.

Tabela 8: Efeitos de ataque estimados através do modelo 2

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.519	0.529	0.149	0.293	0.761
Botafogo	-0.346	-0.356	0.185	-0.613	-0.011
Corinthians	-0.096	-0.081	0.134	-0.324	0.108
Bahia	-0.082	-0.061	0.133	-0.280	0.119
Fluminense	-0.186	-0.211	0.163	-0.442	0.060
Vasco	-0.147	-0.126	0.152	-0.389	0.064
Palmeiras	0.145	0.138	0.168	-0.151	0.424
São Paulo	-0.156	-0.167	0.157	-0.424	0.085
Santos	0.152	0.126	0.145	-0.029	0.411
Atlético-MG	-0.080	-0.048	0.154	-0.338	0.153
Cruzeiro	-0.507	-0.499	0.213	-0.828	-0.165
Grêmio	0.227	0.217	0.153	0.006	0.500
Internacional	-0.099	-0.067	0.166	-0.379	0.159
Goiás	-0.065	-0.030	0.158	-0.359	0.177
Atlético-PR	-0.008	0.002	0.158	-0.315	0.242
Avaí	-0.764	-0.780	0.214	-1.105	-0.406
Chapecoense	-0.393	-0.382	0.235	-0.839	-0.027
CSA	-0.598	-0.604	0.230	-0.942	-0.217
Ceará-SC	-0.231	-0.251	0.177	-0.482	0.039
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 9: Efeitos de defesa estimados através do modelo 2

	Time	Média	Mediana	Desvio-padrão	5%	95%
21	Flamengo	-0.067	-0.025	0.126	-0.315	0.085
22	Botafogo	-0.001	0.001	0.115	-0.208	0.160
23	Corinthians	-0.130	-0.093	0.153	-0.401	0.038
24	Bahia	-0.031	-0.003	0.109	-0.237	0.129
25	Fluminense	0.022	0.005	0.131	-0.182	0.285
26	Vasco	-0.027	-0.020	0.107	-0.216	0.155
27	Palmeiras	-0.181	-0.190	0.146	-0.426	0.017
28	São Paulo	-0.196	-0.162	0.178	-0.513	0.007
29	Santos	-0.146	-0.113	0.150	-0.420	0.028
30	Atlético-MG	0.063	0.021	0.135	-0.122	0.330
31	Cruzeiro	0.031	0.012	0.148	-0.192	0.356
32	Grêmio	-0.061	-0.015	0.121	-0.292	0.098
33	Internacional	-0.067	-0.024	0.119	-0.304	0.080
34	Goiás	0.281	0.286	0.162	0.011	0.524
35	Atlético-PR	-0.161	-0.126	0.159	-0.456	0.021
36	Avaí	0.230	0.231	0.168	-0.017	0.512
37	Chapecoense	0.042	0.010	0.114	-0.115	0.263
38	CSA	0.192	0.175	0.169	-0.028	0.504
39	Ceará-SC	-0.027	-0.010	0.151	-0.278	0.265
40	Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 10: Efeitos de ataque estimados através do modelo 4

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.645	0.650	0.154	0.388	0.884
Botafogo	-0.305	-0.295	0.212	-0.667	0.027
Corinthians	-0.029	-0.025	0.178	-0.324	0.262
Bahia	0.002	0.003	0.177	-0.289	0.285
Fluminense	-0.132	-0.124	0.189	-0.450	0.166
Vasco	-0.132	-0.127	0.190	-0.457	0.158
Palmeiras	0.334	0.333	0.162	0.069	0.604
São Paulo	-0.096	-0.090	0.186	-0.410	0.208
Santos	0.279	0.283	0.168	-0.011	0.549
Atlético-MG	-0.002	0.009	0.188	-0.328	0.292
Cruzeiro	-0.400	-0.392	0.211	-0.770	-0.076
Grêmio	0.352	0.349	0.164	0.089	0.630
Internacional	0.006	0.009	0.176	-0.291	0.293
Goiás	0.049	0.057	0.187	-0.260	0.342
Atlético-PR	0.153	0.153	0.172	-0.129	0.431
Avaí	-0.709	-0.681	0.257	-1.160	-0.330
Chapecoense	-0.302	-0.293	0.202	-0.639	0.006
CSA	-0.522	-0.506	0.246	-0.955	-0.160
Ceará-SC	-0.165	-0.161	0.187	-0.484	0.127
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 11: Efeitos de defesa estimados através do modelo 4

	Time	Média	Mediana	Desvio-padrão	5%	95%
21	Flamengo	-0.114	-0.099	0.156	-0.380	0.122
22	Botafogo	0.010	0.012	0.137	-0.221	0.239
23	Corinthians	-0.144	-0.134	0.144	-0.389	0.065
24	Bahia	-0.014	-0.007	0.135	-0.249	0.198
25	Fluminense	0.034	0.033	0.126	-0.179	0.248
26	Vasco	0.016	0.016	0.135	-0.207	0.230
27	Palmeiras	-0.157	-0.145	0.154	-0.433	0.070
28	São Paulo	-0.194	-0.181	0.153	-0.473	0.030
29	Santos	-0.165	-0.150	0.155	-0.437	0.058
30	Atlético-MG	0.049	0.048	0.134	-0.165	0.268
31	Cruzeiro	0.032	0.029	0.132	-0.184	0.254
32	Grêmio	-0.078	-0.067	0.145	-0.335	0.146
33	Internacional	-0.067	-0.060	0.142	-0.310	0.153
34	Goiás	0.261	0.256	0.146	0.031	0.515
35	Atlético-PR	-0.154	-0.142	0.145	-0.409	0.059
36	Avaí	0.225	0.221	0.139	0.009	0.461
37	Chapecoense	0.104	0.100	0.131	-0.096	0.322
38	CSA	0.162	0.157	0.138	-0.050	0.393
39	Ceará-SC	-0.029	-0.026	0.134	-0.251	0.179
40	Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 12: Efeitos de ataque estimados através do modelo 5

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.648	0.647	0.148	0.401	0.891
Botafogo	-0.287	-0.272	0.198	-0.631	0.008
Corinthians	-0.031	-0.027	0.170	-0.321	0.235
Bahia	0.002	0.010	0.187	-0.321	0.287
Fluminense	-0.170	-0.160	0.202	-0.515	0.147
Vasco	-0.106	-0.094	0.186	-0.416	0.190
Palmeiras	0.333	0.333	0.152	0.090	0.579
São Paulo	-0.078	-0.076	0.170	-0.356	0.205
Santos	0.290	0.294	0.154	0.034	0.536
Atlético-MG	0.030	0.029	0.173	-0.257	0.322
Cruzeiro	-0.394	-0.383	0.197	-0.718	-0.085
Grêmio	0.339	0.346	0.164	0.069	0.603
Internacional	0.012	0.019	0.172	-0.279	0.286
Goiás	0.050	0.058	0.179	-0.248	0.327
Atlético-PR	0.133	0.134	0.162	-0.135	0.387
Avaí	-0.672	-0.657	0.241	-1.095	-0.300
Chapecoense	-0.294	-0.281	0.201	-0.641	0.013
CSA	-0.467	-0.452	0.212	-0.829	-0.138
Ceará-SC	-0.156	-0.150	0.183	-0.463	0.129
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 13: Efeitos de defesa estimados através do modelo 5

	Time	Média	Mediana	Desvio-padrão	5%	95%
21	Flamengo	-0.098	-0.085	0.154	-0.366	0.143
22	Botafogo	0.015	0.013	0.128	-0.194	0.221
23	Corinthians	-0.150	-0.139	0.148	-0.414	0.070
24	Bahia	-0.019	-0.014	0.137	-0.243	0.198
25	Fluminense	0.002	0.006	0.143	-0.229	0.229
26	Vasco	0.016	0.014	0.130	-0.193	0.227
27	Palmeiras	-0.142	-0.133	0.144	-0.395	0.082
28	São Paulo	-0.191	-0.186	0.149	-0.457	0.037
29	Santos	-0.157	-0.144	0.153	-0.429	0.065
30	Atlético-MG	0.063	0.064	0.135	-0.164	0.281
31	Cruzeiro	0.025	0.025	0.133	-0.190	0.242
32	Grêmio	-0.083	-0.071	0.147	-0.343	0.137
33	Internacional	-0.065	-0.058	0.137	-0.299	0.147
34	Goiás	0.260	0.261	0.143	0.037	0.501
35	Atlético-PR	-0.171	-0.161	0.154	-0.435	0.061
36	Avaí	0.217	0.213	0.134	0.005	0.448
37	Chapecoense	0.099	0.098	0.133	-0.113	0.320
38	CSA	0.177	0.172	0.134	-0.036	0.405
39	Ceará-SC	-0.035	-0.028	0.131	-0.257	0.164
40	Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 14: Efeitos de ataque estimados através do modelo 6

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.722	0.717	0.156	0.470	0.982
Botafogo	-0.273	-0.267	0.197	-0.610	0.044
Corinthians	-0.024	-0.020	0.187	-0.342	0.273
Bahia	0.016	0.018	0.193	-0.305	0.331
Fluminense	-0.212	-0.197	0.219	-0.595	0.122
Vasco	-0.094	-0.088	0.199	-0.421	0.219
Palmeiras	0.363	0.364	0.169	0.091	0.643
São Paulo	-0.078	-0.069	0.185	-0.383	0.215
Santos	0.356	0.353	0.170	0.078	0.641
Atlético-MG	0.062	0.070	0.186	-0.254	0.365
Cruzeiro	-0.461	-0.444	0.242	-0.896	-0.098
Grêmio	0.403	0.401	0.167	0.136	0.674
Internacional	0.031	0.036	0.185	-0.279	0.329
Goiás	0.077	0.080	0.179	-0.225	0.371
Atlético-PR	0.129	0.136	0.184	-0.185	0.419
Avaí	-0.708	-0.686	0.260	-1.169	-0.321
Chapecoense	-0.320	-0.310	0.221	-0.691	0.039
CSA	-0.464	-0.459	0.221	-0.840	-0.120
Ceará-SC	-0.158	-0.148	0.183	-0.465	0.128
Fortaleza	0.000	0.000	0.000	0.000	0.000

Tabela 15: Efeitos de defesa estimados através do modelo 6

	Time	Média	Mediana	Desvio-padrão	5%	95%
21	Flamengo	-0.061	-0.054	0.135	-0.296	0.152
22	Botafogo	0.026	0.024	0.139	-0.191	0.248
23	Corinthians	-0.151	-0.138	0.154	-0.412	0.075
24	Bahia	-0.018	-0.015	0.138	-0.248	0.211
25	Fluminense	-0.009	-0.006	0.148	-0.255	0.226
26	Vasco	0.027	0.024	0.133	-0.189	0.251
27	Palmeiras	-0.143	-0.133	0.149	-0.411	0.083
28	São Paulo	-0.201	-0.184	0.160	-0.471	0.034
29	Santos	-0.138	-0.127	0.147	-0.397	0.083
30	Atlético-MG	0.077	0.073	0.141	-0.149	0.307
31	Cruzeiro	0.008	0.012	0.139	-0.229	0.235
32	Grêmio	-0.067	-0.056	0.149	-0.328	0.162
33	Internacional	-0.065	-0.055	0.143	-0.314	0.160
34	Goiás	0.283	0.275	0.152	0.041	0.542
35	Atlético-PR	-0.179	-0.160	0.163	-0.453	0.057
36	Avaí	0.234	0.227	0.145	0.009	0.486
37	Chapecoense	0.092	0.088	0.132	-0.109	0.315
38	CSA	0.188	0.181	0.140	-0.026	0.423
39	Ceará-SC	-0.029	-0.024	0.134	-0.255	0.185
40	Fortaleza	0.000	0.000	0.000	0.000	0.000

6.3 Código Stan dos modelos

Nesta parte, são mostrados os códigos do Stan para os modelos abordados.

Código para o modelo Poisson misto

```
data {
  int<lower=1> G;
  int<lower=1> T;
  int<lower=0, upper=T> h[G];
  int<lower=0, upper=T> a[G];
  int<lower=0> y1[G];
  int<lower=0> y2[G];
}

parameters {
  real home;
  real mu_att;
  real mu_def;
  real<lower=0> sigma_att;
  real<lower=0> sigma_def;
  vector[T] att_raw;
  vector[T] def_raw;
}

transformed parameters {
  vector[T] att;
  vector[T] def;

  for (t in 1:(T-1)) {
    att[t] = att_raw[t];
    def[t] = def_raw[t];
  }

  att[T] = 0;
  def[T] = 0;
}

model {
  for (g in 1:G) {
    y1[g] ~ poisson_log(home + att[h[g]] + def[a[g]]);
    y2[g] ~ poisson_log(att[a[g]] + def[h[g]]);
  }

  for (t in 1:T) {
    att_raw[t] ~ normal(mu_att, sigma_att);
    def_raw[t] ~ normal(mu_def, sigma_def);
  }

  home ~ normal(0, 10);
  mu_att ~ normal(0, 10);
  mu_def ~ normal(0, 10);
  sigma_att ~ cauchy(0, 2.5);
  sigma_def ~ cauchy(0, 2.5);
}

generated quantities {
  vector[G] y1_tilde;
```

```
  vector[G] y2_tilde;
  vector[G] log_lik;

  for (g in 1:G) {
    y1_tilde[g] = poisson_log_rng(home + att[h[g]] + def[a[g]]);
    y2_tilde[g] = poisson_log_rng(att[a[g]] + def[h[g]]);
    log_lik[g] = poisson_log_lpmf(y1[g] | home + att[h[g]] + def[a[g]]) +
    poisson_log_lpmf(y2[g] | att[a[g]] + def[h[g]]);
  }
}
```

Código para o modelo Poisson misto bivariado e suas extensões

```
data {
  int<lower=1> G;
  int<lower=1> T;
  int<lower=0, upper=T> h[G];
  int<lower=0, upper=T> a[G];
  int<lower=0> y1[G];
  int<lower=0> y2[G];
  int<lower=0, upper=1> gamma1;
  int<lower=0, upper=1> gamma2;
}

parameters {
  real mu;
  real home;
  real alpha;
  vector[T] alpha_home;
  vector[T] alpha_away;
  vector[T-1] att_raw;
  vector[T-1] def_raw;
  real<lower=0> sigma_att;
  real<lower=0> sigma_def;
}

transformed parameters {
  vector[T] att;
  vector[T] def;

  for (t in 1:(T-1)) {
    att[t] = att_raw[t];
    def[t] = def_raw[t];
  }

  att[T] = 0;
  def[T] = 0;
}

model {
    vector[G] lambda1;
    vector[G] lambda2;
    vector[G] lambda3;
```

```
    att_raw ~ normal(0, sigma_att);
    def_raw ~ normal(0, sigma_def);
    home ~ normal(0, 10);
    sigma_att ~ cauchy(0, 2.5);
    sigma_def ~ cauchy(0, 2.5);
    mu ~ normal(0, 10);
    alpha ~ normal(0, 1);
    alpha_home ~ normal(0, 1);
    alpha_away ~ normal(0, 1);

    for (g in 1:G) {
      lambda1[g] = exp(mu + home + att[h[g]] + def[a[g]]);
      lambda2[g] = exp(mu + att[a[g]] + def[h[g]]);
      lambda3[g] = exp(alpha + gamma1 * alpha_home[h[g]] + gamma2 * alpha_away[a[g]]);
    }

    y1 ~ poisson(lambda1 + lambda3);
    y2 ~ poisson(lambda2 + lambda3);
}

generated quantities {
  vector[G] lambda1;
  vector[G] lambda2;
  vector[G] lambda3;
  vector[G] y1_tilde;
  vector[G] y2_tilde;
  vector[G] log_lik;

  for (g in 1:G) {
    lambda1[g] = exp(mu + home + att[h[g]] + def[a[g]]);
    lambda2[g] = exp(mu + att[a[g]] + def[h[g]]);
    lambda3[g] = exp(alpha + gamma1 * alpha_home[h[g]] + gamma2 * alpha_away[a[g]]);

    y1_tilde[g] = poisson_rng(lambda1[g] + lambda3[g]);
    y2_tilde[g] = poisson_rng(lambda2[g] + lambda3[g]);

    log_lik[g] = poisson_log_lpmf(y1[g] | mu + home + att[h[g]] +
    def[a[g]] + alpha + gamma1 * alpha_home[h[g]] +
    gamma2 * alpha_away[a[g]]) +
    poisson_log_lpmf(y2[g] | mu + att[a[g]] + def[h[g]] +
    alpha + gamma1 * alpha_home[h[g]] + gamma2 * alpha_away[a[g]]);
  }
}
```

Código para o modelo Poisson misto com mistura

```
data {
  int<lower=1> G;
  int<lower=1> T;
  int<lower=1> C;
  int<lower=0, upper=T> h[G];
  int<lower=0, upper=T> a[G];
  int<lower=0> y1[G];
  int<lower=0> y2[G];
```

```
}

parameters {
  real home;
  real<lower=-3,upper=0> mu_att_raw1;
  real mu_att_raw2;
  real<lower=0, upper=3> mu_att_raw3;
  real<lower=0, upper=3> mu_def_raw1;
  real mu_def_raw2;
  real<lower=-3,upper=0> mu_def_raw3;

  real<lower=0> sigma_att[C];
  real<lower=0> sigma_def[C];
  vector[T] att_raw;
  vector[T] def_raw;
  simplex[3] pi_att[T];
  simplex[3] pi_def[T];
}

transformed parameters {
  vector[T] att;
  vector[T] def;
  vector[C] mu_att;
  vector[C] mu_def;

  for (t in 1:(T-1)) {
    att[t] = att_raw[t];
    def[t] = def_raw[t];
  }
  att[T] = 0;
  def[T] = 0;

  mu_att[1] = mu_att_raw1;
  mu_att[2] = mu_att_raw2;
  mu_att[3] = mu_att_raw3;
  mu_def[1] = mu_def_raw1;
  mu_def[2] = mu_def_raw2;
  mu_def[3] = mu_def_raw3;
}

model {
  real m_att[C];
  real m_def[C];
  vector[G] lambda1;
  vector[G] lambda2;
  vector[G] m_y1;
  vector[G] m_y2;

  home ~ normal(0, 10);
  mu_att_raw1 ~ normal(0, 10) T[-3,0] ;
  mu_att_raw2 ~ normal(0, 0.01);
  mu_att_raw3 ~ normal(0, 10) T[0,3];
  mu_def_raw1 ~ normal(0, 10) T[0,3];
```

```
  mu_def_raw2 ~ normal(0, 0.01);
  mu_def_raw3 ~ normal(0, 10) T[-3,0];

  for (t in 1:T) {
    pi_att[t] ~ dirichlet(rep_vector(1, C));
    pi_def[t] ~ dirichlet(rep_vector(1, C));

    for(c in 1:C) {
      sigma_att[c] ~ cauchy(0, 2.5);
      sigma_def[c] ~ cauchy(0, 2.5);

      m_att[c] = log(pi_att[t, c]) + student_t_lpdf(att[t] | 4, mu_att[c], sigma_att[c]);
      m_def[c] = log(pi_def[t, c]) + student_t_lpdf(def[t] | 4, mu_def[c], sigma_def[c]);
    }

    target += log_sum_exp(m_att) + log_sum_exp(m_def);
  }

  for (g in 1:G) {

    lambda1[g] = home + att[h[g]] + def[a[g]];
    lambda2[g] = att[a[g]] + def[h[g]];

    m_y1[g] = poisson_log_lpmf(y1[g] | lambda1[g]);
    m_y2[g] = poisson_log_lpmf(y2[g] | lambda2[g]);

    target += m_y1[g] + m_y2[g];
  }
}

generated quantities {
  vector[G] y1_tilde;
  vector[G] y2_tilde;
  vector[G] log_lik;

  for (g in 1:G) {
    y1_tilde[g] = poisson_log_rng(home + att[h[g]] + def[a[g]]);
    y2_tilde[g] = poisson_log_rng(att[a[g]] + def[h[g]]);
    log_lik[g] = poisson_log_lpmf(y1[g] | home + att[h[g]] + def[a[g]]) +
    poisson_log_lpmf(y2[g] | att[a[g]] + def[h[g]]);
  }
}
```