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

Mariana de Castro Pasqualini

2022-07-11

Introdução

Desde a década de 80, modelos estatísticos são aplicados em problemas relacionados à esportes. Um exemplo é prever o número de gols marcados por um time em uma partida de futebol, que podem ser modelados como dados de **contagem**.

Introdução

A distribuição de Poisson é uma das mais utilizadas para esse tipo de problema. Serão apresentados 5 modelos propostos na literatura para prever o número de gols nas partidas do Campeonato Brasileiro dos anos 2019.

Os modelos foram implementados no Stan, um software para amostrar modelos bayesianos, juntamente com o R.

Modelo 1

O modelo de efeitos aleatórios proposto por Baio (2010) é definido da seguinte forma:

Seja $\mathbf{y} = (y_{g1}, y_{g2})$ um um vetor de contagens, que são modelados como Poisson independentes condicionais aos parâmetros $\theta = (\theta_{g1}, \theta_{g2})$

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

No qual $j = 1$ representa o time jogando em casa e $j = 2$ indica a equipe visitante. Assumindo um modelo de efeitos aleatórios com função de ligação log, temos:

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

Os índices $h(g)$ representa o time que está jogando em casa no g-ésimo jogo e $a(g)$ o visitante, indo de 1 a $T = 20$.

Priori

A escolha das distribuições a priori do parâmetros se basearam no artigo e foram adaptadas para as parametrizações e particularidades do Stan.

Parâmetros

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

Hiperparâmetros

- $\mu_{att} \sim Normal(0, 10)$
- $\mu_{def} \sim Normal(0, 10)$
- $\sigma_{att} \sim Cauchy(0, 2.5)$
- $\sigma_{def} \sim Cauchy(0, 2.5)$

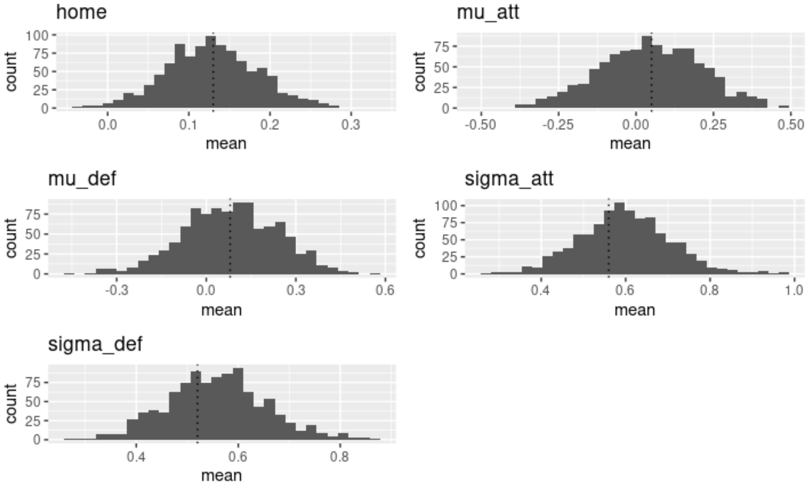
Como **restrição de identificabilidade** nos efeitos específicos de cada time, apenas para o último time foi definido que

$att_{t=20} = 0$ $def_{t=20} = 0$

e, portanto, o vigésimo time é referência para interpretação dos efeitos das outras equipes. Essa restrição foi aplicada para todos os modelos deste trabalho.

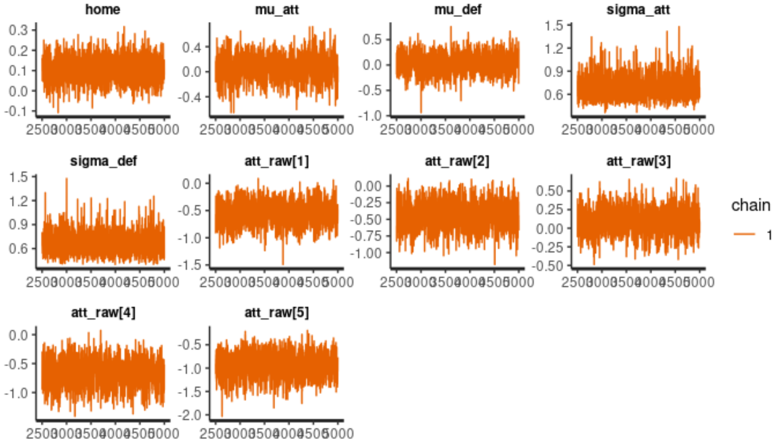
Simulação

Para verificar a implementação e qualidade de estimação dos modelos, foi feita uma simulação com 1000 banco de dados de tamanho 380, representando o número de jogos de um campeonato com 20 times.



Simulação

Uma única cadeia de Markov foi utilizada para obter as amostras da distribuição a posteriori, com 5000 iterações no total, sendo 2500 de warmup/burnin.



Dados

Os dados do Campeonato Brasileiro foram disponibilizados por Gomide e Gualberto no repositório **caRtola**, disponível no Github. O formato dos dados é o seguinte:

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

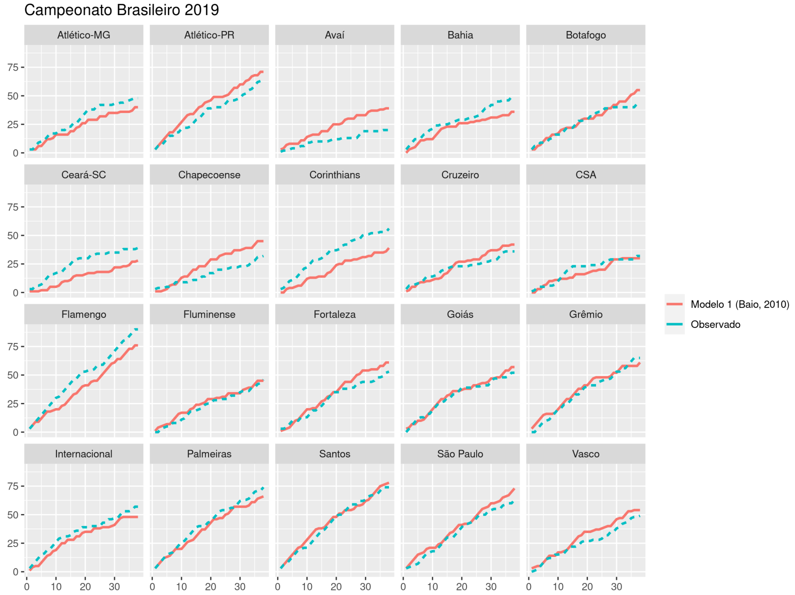
No ano de 2019, o topo da tabela foi formado, respectivamente, por:

- Flamengo, Santos, Palmeiras e Grêmio

Já os times rebaixados foram:

- Cruzeiro, CSA, Chapecoense e Avaí

Ajuste


Campeonato Brasileiro 2019
Atlético-MG
Atlético-PR
Avaí
Bahia
Botafogo
Ceará-SC
Chapecoense
Corinthians
Cruzeiro
CSA
Flamengo
Fluminense
Fortaleza
Goiás
Grêmio
Internacional
Palmeiras
Santos
São Paulo
Vasco
0
25
50
75
0
10
20
30
Modelo 1 (Baio, 2010)
Observado

Ajuste

- Referência: Fortaleza

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.553	0.553	0.152	0.295	0.806
Fluminense	-0.173	-0.166	0.180	-0.474	0.117
Palmeiras	0.227	0.229	0.162	-0.045	0.484
CSA	-0.504	-0.498	0.199	-0.835	-0.187

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	-0.119	-0.116	0.156	-0.384	0.133
Fluminense	-0.032	-0.032	0.148	-0.275	0.209
Palmeiras	-0.194	-0.187	0.162	-0.468	0.054
CSA	0.102	0.096	0.153	-0.147	0.357

Modelo 2

Karlis (2003) propõe um modelo baseado na distribuição de Poisson bivariada, em que são definidas três variáveis aleatórias latentes X_1, X_2, X_3 que seguem, independentemente, uma Poisson com parâmetros $\lambda_1, \lambda_2, \lambda_3$.

Então, as variáveis aleatórias $X = X_1 + X_3$ e $Y = X_2 + X_3$ seguem conjuntamente uma distribuição de Poisson bivariada. Marginalmente, $E(X) = \lambda_1 + \lambda_3$ e $E(Y) = \lambda_2 + \lambda_3$. Ainda, $cov(X, Y) = \lambda_3$ e então tem-se uma medida de interdependência entre as variáveis aleatórias.

Modelo 2

Utilizando o resultado anterior, podemos definir:

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

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

com i indicando o i-ésimo jogo. Daí, temos os preditores lineares definidos como:

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

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

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

no qual as variáveis *dummy* γ_1 e γ_2 indicam quais efeitos queremos incluir na correlação entre o número de gols do time mandante e visitante.

Modelo 2

Para o modelo 2 que estamos considerando, $\gamma_1 = \gamma_2 = 0$. Ou seja, o parâmetro de correlação λ_{3i} depende apenas de um efeito fixo, α.

Originalmente, o artigo considera que o efeito de ataque e defesa para cada time é **fixo**, fazendo com que o modelo tenha um número muito grande de parâmetros. Por isso, o modelo proposto foi adaptado e ataque e defesa foram tratados como efeitos aleatórios, além da inclusão prioris para os parâmetros.

Priori

Parâmetros

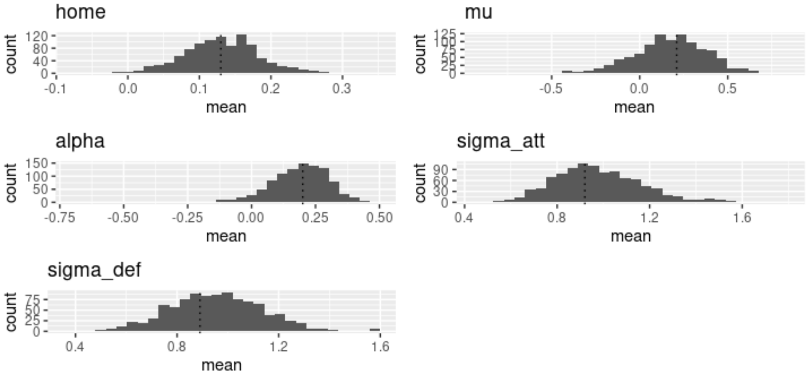
- $\mu \sim \mathit{Normal}(0, 10)$
- $\mathit{home} \sim \mathit{Normal}(0, 10)$
- $\mathit{att}_t \sim \mathit{Normal}(0, \sigma_{\mathit{att}})$
- $\mathit{def}_t \sim \mathit{Normal}(0, \sigma_{\mathit{def}})$
- $\alpha \sim \mathit{Normal}(0, 1)$

Hiperparâmetros

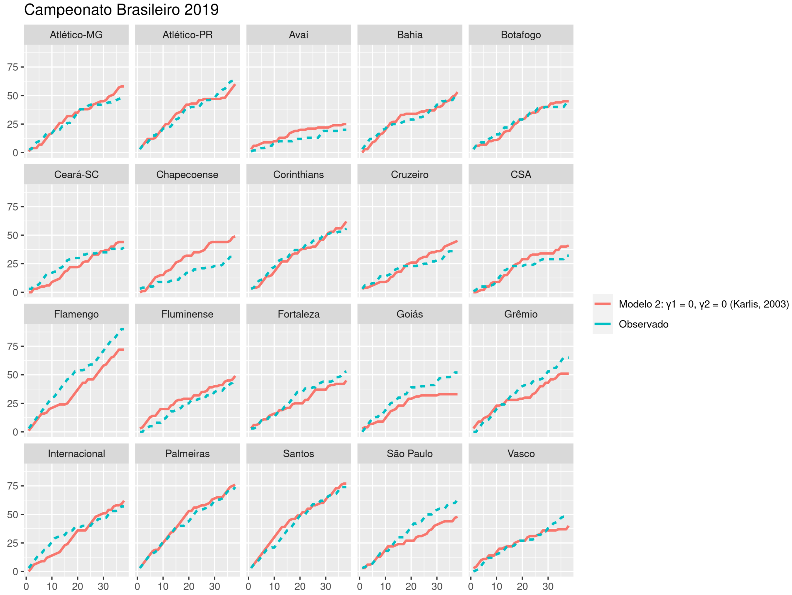
- $\sigma_{\mathit{att}} \sim \mathit{Cauchy}(0, 2.5)$
- $\sigma_{\mathit{def}} \sim \mathit{Cauchy}(0, 2.5)$

Simulação

Uma simulação para o modelo 2 foi feita com 1000 banco de dados de tamanho 380, representando o número de jogos de um campeonato com 20 times.



Ajuste



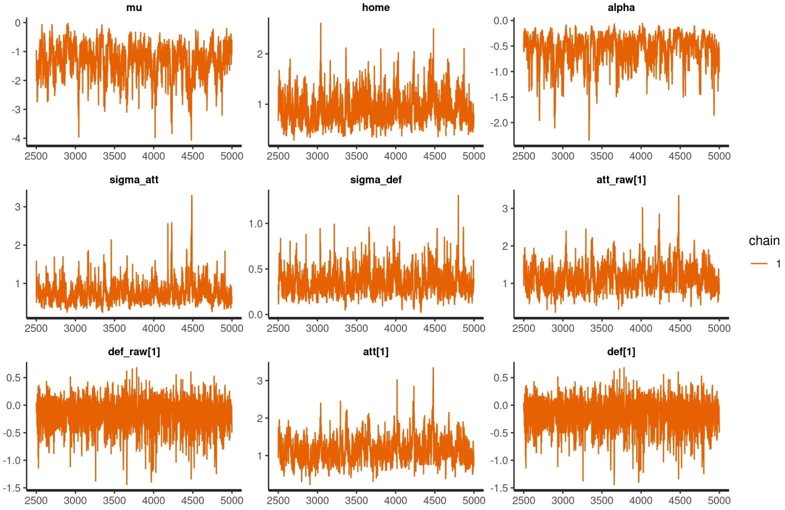
Campeonato Brasileiro 2019
Atlético-MG
Atlético-PR
Avaí
Bahia
Botafogo
Ceará-SC
Chapecoense
Corinthians
Cruzeiro
CSA
Flamengo
Fluminense
Fortaleza
Goiás
Grêmio
Internacional
Palmeiras
Santos
São Paulo
Vasco
Modelo 2: γ1 = 0, γ2 = 0 (Karlis, 2003)
Observado

Ajuste

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	1.164	1.130	0.340	0.697	1.756
Fluminense	-0.521	-0.418	0.581	-1.445	0.169
Palmeiras	0.654	0.616	0.331	0.178	1.233
CSA	-0.830	-0.754	0.567	-1.810	-0.088

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	-0.178	-0.150	0.281	-0.678	0.232
Fluminense	0.052	0.053	0.233	-0.326	0.435
Palmeiras	-0.238	-0.217	0.271	-0.722	0.164
CSA	0.246	0.238	0.240	-0.129	0.645

Ajuste

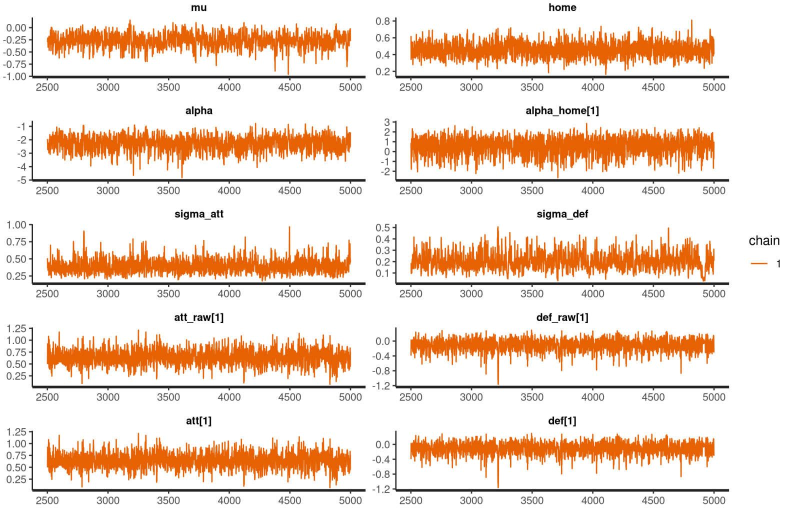


mu
home
alpha
sigma_att
sigma_def
att_raw[1]
def_raw[1]
att[1]
def[1]
chain
1

Modelo 3

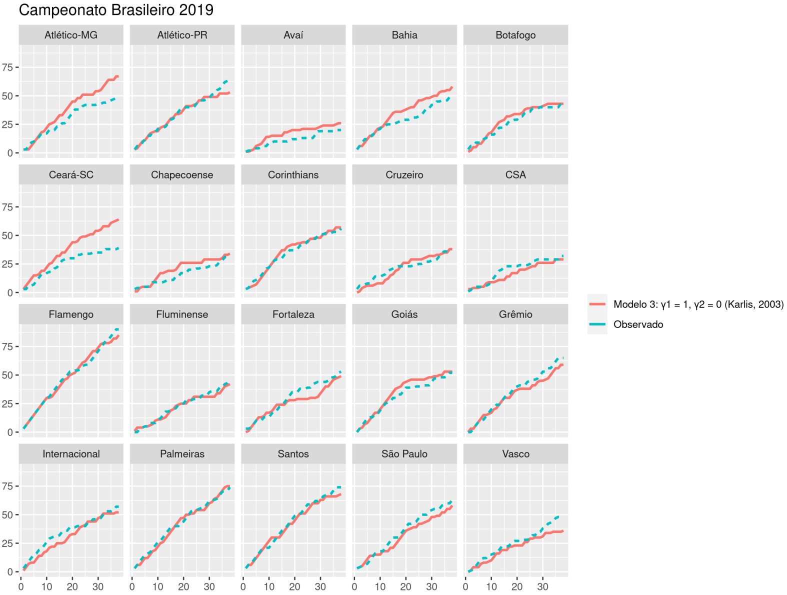
Baseado na definição do modelo 2, o modelo 3 é definido com $\gamma_1 = 1, \gamma_2 = 0$, com parâmetro de correlação λ_{3i} sendo a soma do efeito **fixo** α mais um efeito que depende do time **mandante**.

Modelo 3



mu
home
alpha
alpha_home[1]
sigma_att
sigma_def
att_raw[1]
def_raw[1]
att[1]
def[1]
chain
1

Ajuste

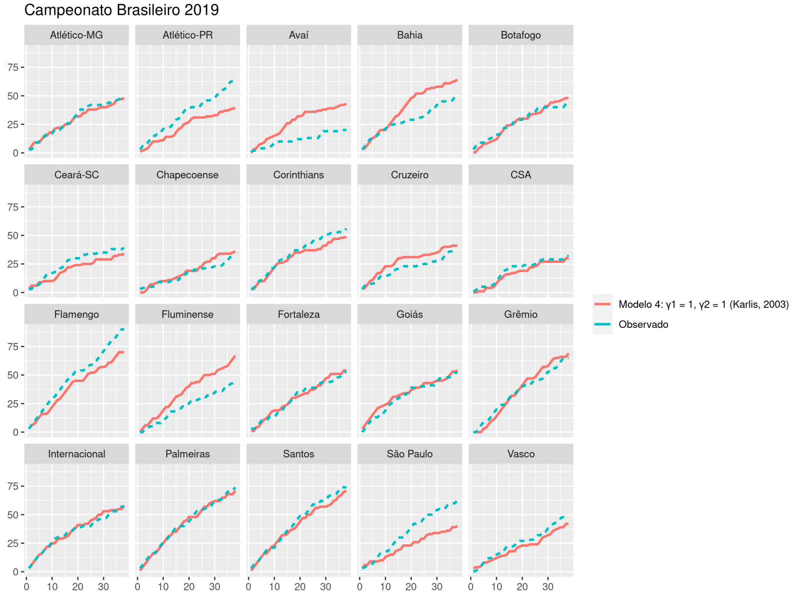


Campeonato Brasileiro 2019
Atlético-MG
Atlético-PR
Avaí
Bahia
Botafogo
Ceará-SC
Chapecoense
Corinthians
Cruzeiro
CSA
Flamengo
Fluminense
Fortaleza
Goiás
Grêmio
Internacional
Palmeiras
Santos
São Paulo
Vasco
0
25
50
75
0
10
20
30
Modelo 3: γ1 = 1, γ2 = 0 (Karlis, 2003)
Observado

Modelo 4

O modelo 4 inclui todos os efeitos na correlação: o efeito **fixo** α, um efeito que depende do time da **casa** e também um efeito para o time **visitante**. Assim, $\gamma_1 = 1, \gamma_2 = 1$.

Ajuste

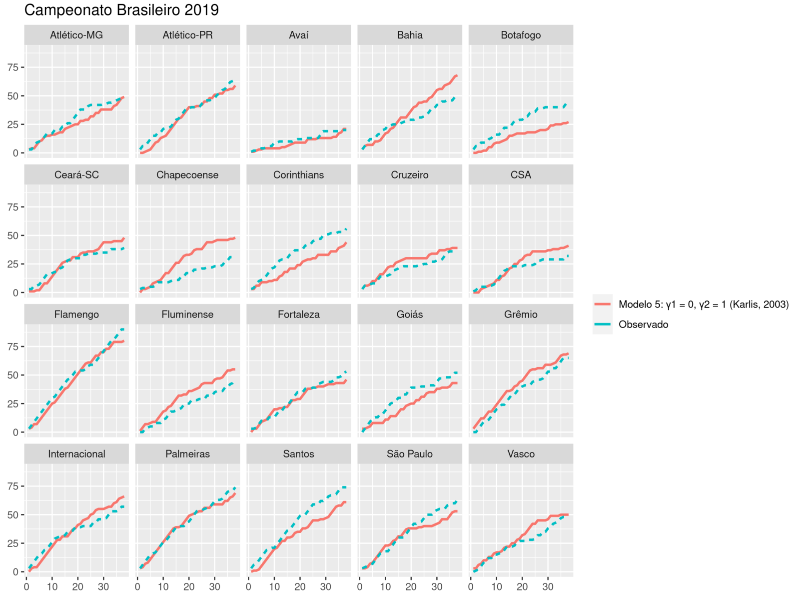


Campeonato Brasileiro 2019
Atlético-MG
Atlético-PR
Avaí
Bahia
Botafogo
Ceará-SC
Chapecoense
Corinthians
Cruzeiro
CSA
Flamengo
Fluminense
Fortaleza
Goiás
Grêmio
Internacional
Palmeiras
Santos
São Paulo
Vasco
0
25
50
75
0
10
20
30
Modelo 4: γ1 = 1, γ2 = 1 (Karlis, 2003)
Observado

Modelo 5

Já o modelo 5 inclui apenas o efeito fixo e o efeito do time **visitante**, definido com $\gamma_1 = 0, \gamma_2 = 1$.

Ajuste



Campeonato Brasileiro 2019
Atlético-MG
Atlético-PR
Avaí
Bahia
Botafogo
Ceará-SC
Chapecoense
Corinthians
Cruzeiro
CSA
Flamengo
Fluminense
Fortaleza
Goiás
Grêmio
Internacional
Palmeiras
Santos
São Paulo
Vasco
0
25
50
75
0
10
20
30
Modelo 5: γ1 = 0, γ2 = 1 (Karlis, 2003)
Observado

Ajuste

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	0.722	0.717	0.156	0.470	0.982
Fluminense	-0.212	-0.197	0.219	-0.595	0.122
Palmeiras	0.363	0.364	0.169	0.091	0.643
CSA	-0.464	-0.459	0.221	-0.840	-0.120

Time	Média	Mediana	Desvio-padrão	5%	95%
Flamengo	-0.061	-0.054	0.135	-0.296	0.152
Fluminense	-0.009	-0.006	0.148	-0.255	0.226
Palmeiras	-0.143	-0.133	0.149	-0.411	0.083
CSA	0.188	0.181	0.140	-0.026	0.423

Modelo 6

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

O ataque e defesa seguem uma distribuição t-Student com 4 graus de liberdade, **ponderados** pela probababilidade 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

- $\pi_{att} \sim Dirichlet([1,1,1])$ e $\pi_{def} \sim Dirichlet([1,1,1])$
- $home \sim Normal(0,10)$
- Para todos os grupos, $\sigma_{att} \sim Cauchy(0,2.5)$ e $\sigma_{def} \sim Cauchy(0,2.5)$

Grupo 1

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

Grupo 2

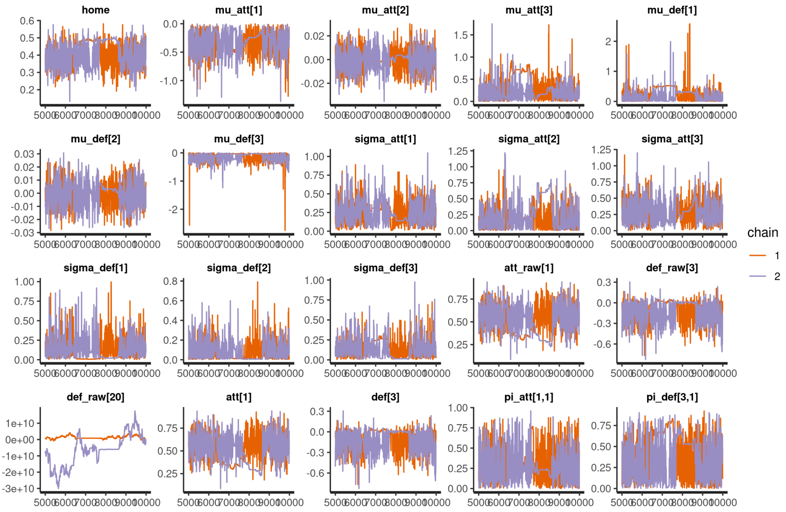
- $\mu_2^{att} \sim truncNormal(0,0.01)$
- $\mu_2^{def} \sim truncNormal(0,0.01)$

Grupo 3

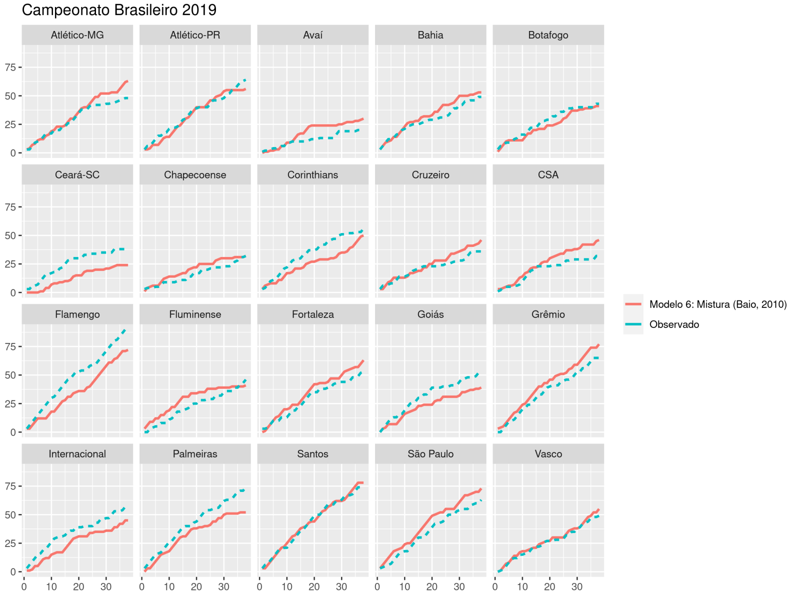
- $\mu_3^{att} \sim truncNormal(0,10,0,3)$
- $\mu_3^{def} \sim truncNormal(0,10,-3,0)$

Modelo 6

As cadeias apresentaram problemas e foram ajustadas com parâmetros diferentes: 2 cadeias, thin = 5 e 10000 iterações.



Ajuste


Campeonato Brasileiro 2019
Atlético-MG
Atlético-PR
Avaí
Bahia
Botafogo
Ceará-SC
Chapecoense
Corinthians
Cruzeiro
CSA
Flamengo
Fluminense
Fortaleza
Goiás
Grêmio
Internacional
Palmeiras
Santos
São Paulo
Vasco
0
25
50
75
0
10
20
30
Modelo 6: Mistura (Baio, 2010)
Observado

Comparação: EQM

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
Avaí	**20**	**39**	**25**	**26**	**43**	**21**	**30**
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
Chapecoense	**32**	**45**	**49**	**34**	**36**	**48**	**32**
Corinthians	56	39	62	57	49	44	50
Cruzeiro	**36**	**42**	**45**	**38**	**41**	**39**	**46**
CSA	**32**	**30**	**41**	**29**	**30**	**41**	**46**
Flamengo	**90**	**76**	**72**	**85**	**70**	**80**	**72**
Fluminense	46	45	49	42	67	55	41
Fortaleza	53	61	45	49	54	46	63
Goiás	52	57	33	53	54	43	39
Grêmio	**65**	**61**	**51**	**59**	**69**	**69**	**77**
Internacional	57	48	62	52	58	66	45
Palmeiras	**74**	**66**	**76**	**75**	**71**	**69**	**52**
Santos	**74**	**78**	**77**	**68**	**71**	**61**	**78**
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

Comparação: LOO-CV

- $elpd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i})$, onde $p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i}d\theta)$ é a densidade preditiva (Vehtari 2015)
- $LOO_{ic} = -2 \times elpd_{loo}$
- Quanto menor, melhor

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

Considerações

- Binomial negativa, para os casos de superdispersão
- Benz (2020) analisa o efeito de jogar em casa durante a pandemia
- Campeonatos de outros anos
- Limitações dos modelos baseados na Poisson bivariada
- Melhoria e correção do modelo de mistura

Referências

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.

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.

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

Vehtari, Andrew e Gabry, Aki e Gelman. 2015. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." https://doi.org/10.48550/ARXIV.1507.04544.