Modelagem e Previsão de Resultados de Partidas de Futebol

Renato Monteiro Pinha Gomes

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Métodos Estatísticos

2018

Modelagem e Previsão de Resultados de Partidas de Futebol

Renato Monteiro Pinha Gomes

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, UFRJ, como parte dos requisitos necessários à obtenção do grau de Mestre em Estatística.

Orientadores: Flávia Maria Pinto Ferreira Landim, João Batista de Morais Pereira.

Rio de Janeiro, RJ - Brasil

2018

Modelagem e Previsão de Resultados de Partidas de Futebol

Renato Monteiro Pinha Gomes

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, UFRJ, como parte dos requisitos necessários à obtenção do grau de Mestre em Estatística.

Aprovada por:

Profª. Flávia M. F. P. Landim, DSc, UFRJ

Prof. Carlos A. Abanto-Valle, DSc, UFRJ

Prof. Leonardo S. Bastos, PhD, Fiocruz

Rio de Janeiro, RJ - Brasil

2018

CIP - Catalogação na Publicação

Gomes, Renato Monteiro Pinha
G633m Modelagem e Previsão de Resultados de Partidas de Futebol / Renato Monteiro Pinha Gomes. -- Rio de Janeiro, 2018.
90 f.

Orientadora: Flávia Maria Pinto Ferreira Landim.
Coorientador: João Batista de Morais Pereira.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Estatística, 2018.

1. Estatística. 2. Futebol. 3. Modelagem. 4. Abordagem bayesiana. I. Landim, Flávia Maria Pinto Ferreira, orient. II. Pereira, João Batista de Morais, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a).

Aos meus pais e irmãos.

"*Clássico é clássico e vice-versa*"

Jardel, jogador.

Agradecimentos

Agradeço a Deus pelo seu infinito amor.

Aos meus pais Celso e Olga que sempre estiveram ao meu lado e que me estimularam a cursar o mestrado.

Aos meus irmãos César, Letícia e Livia pelo carinho e paciência que sempre tiveram comigo.

Ao meu cunhado Pedro, o mais novo integrante da família.

Aos meus tios, em especial ao Silvio Pinha que me ajudou e me estimulou a cursar o curso de Estatística.

Aos meus avôs Rubens e Acyr e às minhas avós Sylvia e Zaira.

Aos meus professores da graduação e do mestrado, em especial: José Francisco, Julio Siqueira, Ricardo Accioly, Eduardo Campos, Maria Elvira, Carlos Abanto-Valle, Maria Eulalia, Flávia Landim e Fernando Moura.

Aos meus orientadores Flávia e João pelo auxilio fornecido para elaboração da minha dissertação. Minha eterna gratidão pelo auxílio e orientações dadas.

A todos os meus amigos, em especial: Allan, Ayrton, Daniela, Gabriel, Humberto, Isabela, Luiz, Marcel, Marcus, Pedro, Rafael, Raíra, Rebecca, Roberta, Rodrigo, Victor Archanjo, Victor Eduardo e Wagner.

À CAPES e à FAPERJ pelo apoio financeiro dos meus estudos.

Por fim aos professores Carlos Abanto-Valle e Leonardo Bastos que aceitaram fazer parte da banca.

Resumo

No cenário esportivo, técnicas estatísticas estão sendo cada vez mais utilizadas com finalidades diversas, como fornecer informações para melhorar o desempenho das equipes na avaliação de jogadores e na previsão de resultados. Percebe-se que vários veículos de comunicação utilizam essas ferramentas para exibir dados ligados a esportes como, por exemplo, preferência do público com relação à determinada modalidade esportiva, média de público e renda, entre outros.

O foco da dissertação é estudar e desenvolver modelos de previsão para resultados das partidas de futebol. Modelos serão propostos para previsão dos placares em que assume-se fatores latentes para explicar ataque, defesa e efeito do mando de campo das equipes.

Considerou-se diferentes modelos: assumindo que os fatores são estáticos ao longo das rodadas; que eles evoluem no tempo de forma dinâmica; que eles evoluem no tempo por meio de componentes auto-regressivas; e assumindo estruturas hierárquicas de regressão.

O procedimento de inferência será feito sob o enfoque bayesiano. Como as distribuições *a posteriori* não são analiticamente tratáveis, adotou-se o Método de Monte Carlo via Cadeias de Markov (MCMC), em particular utilizando os algoritmos amostrador de Gibbs e Metropolis-Hastings para obter amostras dessa distribuição.
Palavras-Chave: estatística; futebol; modelagem; abordagem bayesiana.

Abstract

In sports scene, statistical techniques are being increasily used for several purposes, such as providing information to improve teams' performance at evaluation of players, and prediction of results. Various communication vehicles use these tools to display data related to sports such as, for example, public preference concerning to a particular sport modality, average audience and income, among others.

The focus of dissertation is to study and develop predictive models from results to soccer matches. Models are proposed for prediction of scoreboards which latent factors assumed to explain attack, defense and the effect from teams' field command.

Different models were considered: assuming that factors are static along the matches; they evolve dynamically over time by means of autoregressive components; and assuming hierarchical regression structures.

The inference procedure is done under the bayesian approach. The *posteriori* distribution is not analytically tractable, then Monte Carlo's method via Markov Chains (MCMC) was adopted, in particular using the algorithms Sampler of Gibbs and Metropolis-Hastings to obtain samples from it.
Keywords: Keywords: statistic; soccer; modeling; Bayesian approach.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Considerações gerais

No cenário esportivo, técnicas estatísticas estão sendo cada vez mais utilizadas com finalidades diversas, como fornecer informações para melhorar o desempenho das equipes na avaliação de jogadores e na previsão de resultados. Percebe-se que vários veículos de comunicação utilizam essas ferramentas para exibir dados ligados a esportes como, por exemplo, preferência do público com relação à determinada modalidade esportiva, média de público e renda, entre outros.

No que tange à aplicação de modelos estatísticos na previsão de resultados nas partidas de futebol, a literatura disponibiliza uma variedade de modelos (Dixon e Coles (1997); Rue e Salvesen (2000); Souza Júnior e Gamerman (2004); Louzada et al. (2015)). Pode-se dizer que o futebol é o esporte mais popular no Brasil. Diferente de outros esportes, como basquete e vôlei, uma característica importante do futebol é a grande incerteza nos resultados das partidas realizadas entre as equipes. Muitas vezes times com grande investimento financeiro perdem para clubes com baixo investimento, algo mais difícil de acontecer no basquete, por exemplo. Essa é uma das características que o torna apaixonante e que vem despertando o interesse de vários pesquisadores visando a criação e implementação de modelos capazes de prever resultados e avaliar as equipes no decorrer de um campeonato.

No Brasil o campeonato de futebol de maior destaque é o Campeonato Brasileiro

de Futebol Série A. A edição 2017 foi disputada num sistema de pontos corridos, com jogos de ida e volta. As 20 equipes participantes jogarão em grupo único, todas contra todas. A equipe que marcar mais pontos ao final das 38 rodadas será a campeã. Se uma ou mais equipes terminarem com o mesmo número de pontos, os critérios de desempate definirão as posições. Os seis primeiros colocados foram classificados para a disputa da Copa Libertadores da América de 2018. Os quatro últimos foram rebaixados para a disputa do Campeonato Brasileiro Série B em 2018.

Muitas vezes pesquisadores conhecem informações que impactaram ou ainda podem impactar uma partida de futebol. A modelagem bayesiana permite que tais informações externas sejam incorporadas nos modelos tanto no processo de estimação quanto no de previsão, possibilitando que os usuários possam fazer intervenções subjetivas. Pelo teorema de Bayes associa-se essas informações *a priori* dos pesquisadores e os dados obtidos na amostra. Nesse estudo, toda a abordagem será feita sob o paradigma bayesiano.

1.2 Objetivo da dissertação

O foco da dissertação é estudar e desenvolver modelos de previsão para resultados das partidas de futebol utilizando a abordagem bayesiana. Modelos serão propostos para previsão dos placares em que assume-se fatores latentes para explicar ataque, defesa e efeito do mando de campo das equipes.

Considerou-se diferentes modelos: assumindo que os fatores são estáticos ao longo das rodadas; que eles evoluem no tempo de forma dinâmica; que eles evoluem no tempo por meio de componentes auto-regressivas; e assumindo estruturas hierárquicas de regressão.

Espera-se que os novos modelos propostos nessa dissertação contribuam para a área de estudo esportiva visando uma melhor capacidade de previsão de resultados de partidas de futebol em campeonatos de pontos corridos.

Os dados utilizados nos modelos são do Campeonato Brasileiro de Futebol Série A edição 2017. No entanto, os mesmos modelos poderiam ser aplicados a dados de outros campeonatos com sistemas de pontos corridos, como por exemplo o Campeonato Brasileiro de Futebol Série B entre outros.

1.3 História do futebol

Será apresentado um breve resumo sobre a história do futebol e de como o esporte chegou no Brasil. Maiores detalhes sobre a história do futebol podem ser encontrados em Poli e Carmona (2009) e de como o esporte chegou ao Brasil em Gambeta (2015).

O futebol foi criado no dia 23 de outubro de 1863 na Inglaterra, quando representantes de onze escolas se reuniram com o objetivo de estabelecer regras comuns, visto que cada escola possuía regras distintas. Nesse dia foi fundada a primeira entidade dirigente do futebol mundial, a *Football Association*, sob a direção de Ebenezer Cobb Morley.

No início do século XX, o esporte já tinha se espelhado em outros países. Em 1904, reuniram-se em Paris sete associações dos países da França, Bélgica, Dinamarca, Holanda, Espanha, Suécia e Suíça para fundar a *Fédération Internationale de Football Association* (FIFA). A Inglaterra veio a se associar mais tarde no ano de 1906. Depois de mais de cem anos, o futebol tornou-se o esporte mais popular do mundo. Duzentos e onze federações são associadas a FIFA, chegando a ser apelidada de ONU no futebol.

O futebol chegou ao Brasil em 1894, através do inglês Charles Miller. Embora exista uma corrente que contesta a chegada do esporte afirmando que antes de Miller já existia a prática do futebol, a maioria dos especialistas e historiadores concordam que a organização de times e a adoção das regras oficiais foram implementadas por Miller.

Duas décadas depois foi criada a Federação Brasileira de Esportes, futura Confederação Brasileira de Futebol (CBF). Atualmente, o campeonato de maior destaque realizado no Brasil é o Campeonato Brasileiro de Futebol Série A, organizado pela CBF.

1.4 Futebol

No futebol duas equipes compostas por onze jogadores cada se enfrentam em um campo retangular, sendo supervisionadas por um árbitro. Em cada um dos dois lados menores do retângulo, também conhecidas como linhas de fundo, existe uma baliza. O objetivo do esporte é deslocar a bola pelo campo para colocá-la dentro da baliza adversária. Tal ação é denominada gol. A equipe que fizer o maior número de gols vence

a partida. Cada partida é composta por dois tempos de 45 minutos e um intervalo de 15 minutos entre os tempos. Vale destacar que, exceto os goleiros, que são responsáveis em defender a baliza dos seus respectivos times, todos os outros jogadores não podem colocar a mão na bola.

O campo é composto pelos seguintes elementos: pequena área (espaço onde se realiza a cobrança do tiro de meta); grande área (determina o espaço onde o goleiro pode usar as mãos); linha lateral e linha de fundo (delimita o espaço do campo); círculo central (delimita o espaço do toque inicial da bola); ponto central (marca no centro do círculo central onde a bola deve ficar para receber o primeiro toque); linha de meio de campo (divide o campo ao meio e delimita as áreas de cada equipe), arco-penal (meia circunferência que determina a distância que os jogadores não envolvidos em penalidades devem permanecer durante a cobrança dos pênaltis) e tiro penal (determina o local para cobrança dos pênaltis).

A cobrança de escanteio é marcada quando a bola sai pela linha de fundo e toca por último em algum jogador da equipe que estava se defendendo. O tiro de meta é marcado caso a bola toque por último em algum jogador da equipe que estava atacando e saia pela linha de fundo da equipe adversária. A cobrança de lateral é sinalizada quando a bola sai pela linha lateral, sendo marcada contra a equipe do último jogador que tocou na bola antes de sair pela linha lateral.

Quando um jogador comete faltas ou algum ato de indisciplina, ele pode ser punido com um cartão amarelo ou um vermelho pelo árbitro. Caso algum jogador seja punido por dois cartões amarelos ou por um vermelho em uma partida, ele é expulso do jogo e sua equipe fica com um jogador a menos. Além disso, se algum jogador cometer uma falta na sua grande área, é marcada a cobrança de pênalti para a equipe adversária.

Para evitar que os jogadores fiquem na área da equipe adversária, foi criado a regra do impedimento, que impede o lance caso no momento em que a bola tenha sido tocado para o jogador não tiver pelo menos dois jogadores da equipe adversária entre ele e a linha de fundo adversária. Desse modo, dois árbitros assistentes, conhecidos como bandeirinhas, ficam um em cada linha lateral controlando os impedimentos e também auxiliando o árbitro em marcações de faltas.

1.5 Campeonato Brasileiro de Futebol

O Campeonato Brasileiro de Futebol começou no ano de 1971, sendo campeão o clube Atlético-MG. Antes dele, existiu a Taça Brasil (1959 a 1969), o Torneio Roberto Gomes Pedrosa (1967 a 1970), entre outros. Recentemente a CBF unificou os títulos brasileiros em que foram incorporados os times campeões da Taça Brasil e o Torneio Roberto Gomes Pedrosa. Desse modo oficialmente o primeiro campeão passou a ser o Bahia-BA em 1959. Sendo assim, uma vez unificados os títulos, os clubes com maiores quantidades de títulos são: Palmeiras-SP (9 títulos), Santos-SP (8 títulos) e Corinthians-SP (7 títulos).

Durante muitos anos a estrutura do campeonato era alterada frequentemente. Regulamento, número de times e inclusive o nome do torneio foram alterados ao longo das edições. No ano de 2003 ocorreu uma mudança importante. O campeonato passou a ser disputado num sistema de pontos corridos, com jogos de ida e volta. Os 24 participantes jogaram em grupo único, todos contra todos. O clube que marcou mais pontos ao final das 46 rodadas foi declarado Campeão Brasileiro de 2003. O campeão, o vice, o terceiro e o quarto colocados foram classificados para a disputa da Copa Libertadores da América edição 2004. Os dois últimos times foram rebaixados para a disputa da série B em 2004. O campeão da série B e o vice foram automaticamente promovidos para a disputa da série A em 2004.

Com essa nova configuração, o campeonato ficou mais organizado tanto para os clubes quanto para os seus torcedores, tornando o torneio mais competitivo e mais atrativo para investimentos. Poucas mudanças foram realizadas nas temporadas posteriores, no entanto, a alteração do número de clubes teve grande destaque, passando para vinte no ano de 2006.

Na temporada de 2017 da série A, o campeonato é composto por 38 rodadas, com dez partidas em cada, totalizando 380 partidas. Os vinte clubes que estavam na disputa foram: Atlético-GO (ACG), Atlético-MG (CAM), Atlético-PR (CAP), Avaí-SC (AVA), Bahia-BA (BAH), Botafogo-RJ (BOT), Chapecoense-SC (CHA), Corinthians-SP (COR), Coritiba-PR (CFC), Cruzeiro-MG (CRU), Flamengo-RJ (FLA), Fluminense-RJ (FLU), Grêmio-RS (GRE), Palmeiras-SP (PAL), Ponte Preta-SP (PON), Santos-SP (SAN), São

Paulo-SP (SAO), Sport-PE (SPO), Vasco da Gama-RJ (VAS) e Vitória-BA (VIT). É importante destacar que uma vitória representa o ganho de três pontos, empate, um, e derrota, nenhum. O clube que conquistar a maior quantidade de pontos será o campeão e os últimos quatro serão rebaixados para a série B. Em caso de empate, serão adotados os seguintes critérios, nessa ordem: maior número de vitórias, maior saldo de gols, maior número de gols pró, confronto direto, menor número de cartões vermelhos, menor número de cartões amarelos e sorteio.

Decidiu-se utilizar para esse estudo os dados referentes ao Campeonato Brasileiro da série A edição 2017 até a trigésima quinta rodada, por uma questão de tempo para a conclusão da dissertação, uma vez que o campeonato terminava no início de dezembro. Os dados foram coletados no site *Soccerway* (disponível em https://br.soccerway.com/national/brazil/serie-a/2017). Em todas as partidas realizadas considerou-se um time mandante e um visitante, ou seja, não foram consideradas realizações de partidas em campo neutro.

1.6 Estrutura do texto

O presente trabalho está organizado em seis capítulos sendo o primeiro a Introdução. A seguir, no Capítulo 2 é apresentada uma revisão sobre inferência estatística. Na sequência, no Capítulo 3, são apresentadas noções básicas de modelos lineares generalizados, modelos lineares dinâmicos e modelos lineares dinâmicos generalizados com foco na distribuição de Poisson. Os modelos ajustados, tanto os propostos na literatura quando os propostos nesta dissertação, são descritos no Capítulo 4. Os resultados obtidos são apresentados no Capítulo 5. Por fim, no Capítulo 6 são apresentadas as conclusões do trabalho.

Capítulo 2

Inferência estatística

Em muitas situações pesquisadores querem descrever determinado fenômeno através de uma lei ou modelo de probabilidade. Para isso, utilizando as técnicas adequadas, retira-se uma amostra da população e, de posse desses dados, deseja-se descrever e fazer inferência com os valores sobre esta população. A inferência estatística é o conjunto de técnicas que visa através de informações obtidas a partir da amostra descrever e analisar determinado fenômeno aleatório em uma população.

Geralmente no processo de inferência tem-se dois tipos de informação: a informação *a priori*, ou seja, anterior ou externa ao processo de amostragem, advinda de conhecimentos do pesquisador ou da própria natureza do problema investigado, e a informação obtida a partir da amostra. Nesse contexto pode-se citar dois tipos de abordagens na inferência estatística: a clássica (ou frequentista) e a bayesiana. Em ambas modelos probabilísticos são assumidos para descrever o fenômeno de interesse cujos parâmetros são quantidades desconhecidas a serem estimadas. O tratamento e a interpretação em relação aos parâmetros é o diferencial das duas abordagens. Na clássica o parâmetro é um valor desconhecido porém fixo, empregando-se exclusivamente a informação obtida por amostragem para sua estimação. Na bayesiana assume-se que a incerteza à respeito de um parâmetro é caracterizada por uma distribuição *a priori*, que quando combinada com a informação da amostra, dá origem a distribuição *a posteriori*, na qual se baseia o procedimento de inferência.

Nesse capítulo será descrita de maneira breve a abordagem bayesiana, uma vez que

nesse trabalho optou-se por utilizar tal abordagem nos modelos que serão apresentados posteriormente. Para uma visão mais profunda e detalhada sobre inferência estatística, ver Casella e Berger (2010); DeGroot e Schervish (2012); Migon (2015); Robert e Casella (2004).

2.1 Abordagem bayesiana

Como dito anteriormente na abordagem bayesiana, a incerteza a respeito de um parâmetro ou vetor paramétrico é caracterizada por uma distribuição a priori. Uma vez realizado o processo amostral, através do teorema de Bayes, combina-se a distribuição *a priori* com a função de verossimilhança obtendo-se a distribuição *a posteriori*, que é dada por:

$$P(\boldsymbol{\theta}|\mathbf{Y}) = \frac{P(\boldsymbol{\theta}, \mathbf{Y})}{P(\mathbf{Y})} = \frac{P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta})}{\sum P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.1}$$

no caso discreto, e

$$P(\boldsymbol{\theta}|\mathbf{Y}) = \frac{P(\boldsymbol{\theta}, \mathbf{Y})}{P(\mathbf{Y})} = \frac{P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta})}{\int P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.2}$$

no caso contínuo. O denominador das expressões acima não dependem de $\boldsymbol{\theta}$. Sendo assim o denominador é apenas uma constante. Pode-se então reescrever as equações apresentadas anteriormente da seguinte forma:

$$P(\boldsymbol{\theta}|\mathbf{Y}) \propto P(\boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta}) \tag{2.3}$$

Essa última apresentação da distribuição *a posteriori* retrata bem a combinação de informações *a priori* com a amostra obtida. Uma vez identificado o núcleo da distribuição *a posteriori* pode-se obter estimativas pontuais ou intervalares para os parâmetros. Em muitas ocasiões não é possível obter uma expressão analítica fechada para distribuição *a posteriori*. Nesses casos, entre diferentes abordagens, pode-se utilizar métodos de simulação estocástica para a obtenção de uma amostra da distribuição *a posteriori*.

Outra forma de expressar a distribuição *a posteriori* corresponde a atualizá-la sequencialmente cada vez que uma nova observação y_i (condicionalmente independentes entre si) é obtido, ou seja:

$$\begin{aligned}
P(\boldsymbol{\theta}|y_1) &\propto P(\boldsymbol{\theta})P(y_1|\boldsymbol{\theta}) \\
P(\boldsymbol{\theta}|y_1, y_2) &\propto P(\boldsymbol{\theta}|y_1)P(y_2|\boldsymbol{\theta}) \\
&\propto P(\boldsymbol{\theta})P(y_1|\boldsymbol{\theta})P(y_2|\boldsymbol{\theta}) \\
&\vdots \\
P(\boldsymbol{\theta}|y_1, y_2, ..., y_n) &\propto P(\boldsymbol{\theta}|y_1, y_2, ..., y_{n-1})P(y_n|\boldsymbol{\theta}) \\
&\propto P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(y_i|\boldsymbol{\theta}).
\end{aligned}$$

Diz-se que a distribuição *a priori* é conjugada quando a distribuição *a posteriori* pertence a mesma classe da distribuição *a priori*. Alguns exemplos de distribuições conjugadas a determinado processo de amostragem são: distribuição beta conjugada ao modelo Binomial; distribuição gama conjugada ao modelo Poisson e a distribuição normal conjugada ao modelo normal. Além disso diz-se que a distribuição *a priori* é vaga se ela traz pouca ou nenhuma informação, tendo pouca contribuição na informação *a posteriori*.

Geralmente não é fácil a escolha das distribuições *a priori* para os parâmetros. Em alguns casos utiliza-se distribuições *a prioris* conjugadas para facilitar a obtenção de distribuições *a posteriori* conhecidas. Entretanto, algumas vezes tais distribuições não refletem com fidelidade o conhecimento prévio a respeito do parâmetro. Um recurso muito utilizado quando não se tem informações a respeito dos parâmetros é a atribuição de distribuições *a prioris* vagas.

A partir da distribuição *a posteriori*, pode-se obter a distribuição preditiva de $Y_{n+1}|\mathbf{Y}$, que é dada por:

$$
\begin{aligned}
P(y_{n+1}|\mathbf{Y}) &= \int P(y_{n+1}, \boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} \\
&= \int P(y_{n+1}|\boldsymbol{\theta}, \mathbf{Y}) P(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} \\
&= \int P(y_{n+1}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}.
\end{aligned}
$$

Nessa última passagem supõe-se que, condicionada a $\boldsymbol{\theta}$, Y_{n+1} e $\mathbf{Y}$ são independentes.

2.1.1 Estimadores pontuais

Considere uma distribuição *a posteriori* $\boldsymbol{\theta}|\mathbf{Y}$. Seja Θ o espaço paramétrico, Λ o conjunto de decisões possíveis e $a \in \Lambda$ uma ação. A função perda $L(\boldsymbol{\theta}, a)$ é uma função $L : \Theta \times \Lambda \to [0, +\infty)$ interpretada como a perda sofrida ao estimar $\boldsymbol{\theta}$ por a. Define-se a perda esperada *a posteriori* como:

$$
E[L(\boldsymbol{\theta}, a)|\mathbf{Y}] = \int L(\boldsymbol{\theta}, a) \pi(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} \tag{2.4}
$$

O estimador pontual bayesiano é obtido minimizando a perda esperada *a posteriori*. Existem na literatura muitas funções perdas que podem ser utilizadas. Para cada uma tem-se o estimador pontual para θ. Abaixo serão apresentadas as funções perdas mais aplicadas:

$$
\begin{aligned}
\text{Função perda quadrática: } L(\boldsymbol{\theta}, a) &= (\boldsymbol{\theta} - a)^2 ; \\
\text{Função perda absoluta: } L(\boldsymbol{\theta}, a) &= |\boldsymbol{\theta} - a|; \\
\text{Função perda 0-1: } L(\boldsymbol{\theta}, a) &= 0, \text{ se } |\boldsymbol{\theta} - a| < \epsilon \text{ ou } 1, \text{ se } |\boldsymbol{\theta} - a| > \epsilon; \text{ para } \epsilon > 0.
\end{aligned}
$$

Os estimadores para a funções perda quadrática, perda absoluta e perda 0-1 são a média, mediana e moda *posteriori* respectivamente.

2.1.2 Estimadores intervalares

Assim como na abordagem clássica na bayesiana também pode-se obter estimadores intervalares para $\boldsymbol{\theta}$. Mas agora não é necessária fazer a interpretação frequentista apre-

sentada anteriormente. Os intervalos, agora chamados intervalos de credibilidade, são calculados de maneira natural através da distribuição *a posteriori*. Uma região $C \in \boldsymbol{\Theta}$ é um intervalo de credibilidade $100(1-\alpha)\%$ para $\boldsymbol{\theta}$ se:

$$P(\boldsymbol{\theta} \in C|\mathbf{Y}) \geq 1-\alpha. \tag{2.5}$$

Agora $1-\alpha$ é chamado nível de credibilidade. No caso escalar, a região C é usualmente dada pelo intervalo $[c_1, c_2]$. O comprimento do intervalo traz informações com relação a concentração da distribuição *a posteriori*. Note que quanto maior for o comprimento do intervalo mais dispersa está a distribuição desse parâmetro e quanto menor for menos dispersa está a distribuição. Além disso, a exigência de que a probabilidade seja maior do que o nível de credibilidade é meramente técnica, pois deseja-se que o intervalo tenha menor comprimento possível, o que em geral implica usar a igualdade. Nos casos em que a distribuição $\boldsymbol{\theta}|\mathbf{Y}$ é discreta a desigualdade é útil visto que nem sempre pode-se obter a igualdade.

2.2 Métodos de simulação via cadeias de Markov

Em muitas ocasiões não é possível ou é muito complexo a obtenção da expressão fechada da distribuição *a posteriori*. Nesses ocasiões pode-se aplicar métodos de simulação para obtenção de uma ou mais amostras da distribuição. Os métodos apresentados nessa subseção são conhecidos como métodos de Monte Carlo via cadeias de Markov (mais detalhes podem ser vistos em Gamerman e Lopes (2006)).

A ideia central dos métodos é a construção de uma cadeia de Markov, cuja distribuição estacionária seja coincida com a distribuição de interesse, ou seja, a distribuição *a posteriori*. Valores são simulados iterativamente desta cadeia de até que a convergência seja atingida, ou seja, quando os valores sorteados são assumidos gerados da distribuição estacionária. Maiores detalhes sobre os algoritmos que serão apresentados podem ser encontrados em Gamerman e Lopes (2006) e Robert e Casella (2004).

A verificação da convergência foi feita de forma gráfica por meio da análise dos traços das cadeias dos parâmetros de interesse, iniciadas em diferentes valores.

2.2.1 Metropolis-Hastings

O algoritmo de Metropolis-Hastings (Metropolis et al., 1953) consiste em gerar um valor a partir de uma distribuição auxiliar proposta, que será aceito como um novo valor da cadeia com uma probabilidade dada. Considere que na iteração t a cadeia esteja no estado $\boldsymbol{\theta}^t$. Um valor $\boldsymbol{\theta}'$ é gerado de uma distribuição proposta $q(.|\boldsymbol{\theta}^t)$. O novo valor gerado é aceito com probabilidade:

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}') = min\left(1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^t|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^t)q(\boldsymbol{\theta}'|\boldsymbol{\theta}^t)}\right).$$

A cadeia vai passar para o novo estado $\boldsymbol{\theta}'$ na iteração $t+1$ caso seja aceito. Se rejeitado, o que acontece com probabilidade $1-\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}')$, permanece no estado $\boldsymbol{\theta}^t$. Tal algoritmo pode ser ilustrado pelos seguintes passos:

(1) Inicie o contador $t = 0$.
(2) Informe um valor inicial $\boldsymbol{\theta}^0$.
(3) Determine o número de iterações para geração da cadeia.
(4) Gere um valor $\boldsymbol{\theta}'$ dessa distribuição $q(.|\boldsymbol{\theta}^t)$ proposta.
(5) Calcule $\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}')$.
(6) Gere $u \sim U[0, 1]$.
(7) Se $u \leqslant \alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}')$, aceite o novo valor e faça $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}'$, caso contrário, rejeite e faça $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$.
(8) Incremente o contador de t para $t+1$.
(9) Repita os passos de 4 a 8 até que a convergência seja obtida e até obter o tamanho da amostra necessário.

2.2.2 Amostrador de Gibbs

No amostrador de Gibbs (Geman e Geman, 1984) as probabilidades de transição dos estados são geradas a partir das distribuições condicionais completas. Suponha $p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, ..., \theta_n|\mathbf{Y})$ e considere $p(\theta_i|\mathbf{Y}, \theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, \theta_{i+2}, ..., \theta_n)$, ou seja, a distribuição condicional completa de θ_i.

O algoritmo pode ser descrito pelos seguintes passos:

(1) Inicie o contador $t = 0$.

(2) Informe um valor inicial $\boldsymbol{\theta}^0$.

(3) Gere valores das distribuições condicionais completas.

$$\begin{aligned}
\theta_1^t &\sim P(\theta_1|\mathbf{Y}, \theta_2^{t-1}, \theta_3^{t-1}, \theta_4^{t-1}, ..., \theta_n^{t-1}) \\
\theta_2^t &\sim P(\theta_2|\mathbf{Y}, \theta_1^{t}, \theta_3^{t-1}, \theta_4^{t-1}, ..., \theta_n^{t-1}) \\
&\vdots \\
\theta_n^t &\sim P(\theta_n|\mathbf{Y}, \theta_1^{t}, \theta_2^{t}, \theta_3^{t}, \theta_4^{t}, ..., \theta_{n-1}^{t})
\end{aligned}$$

(4) Faça $t = t + 1$,

(5) Repita os passos de 3 e 4 até obter a convergência e o tamanho de amostra desejado.

Após a convergência, todos os valores obtidos formam a amostra para distribuição *a posteriori*. Observe que as probabilidades de aceitação desse amostrador são iguais a 1, ou seja, a cadeia sempre se moverá.

Capítulo 3

Modelos lineares generalizados e modelos lineares dinâmicos generalizados

3.1 Modelos lineares generalizados (MLG)

Um modelo linear generalizado (MLG) (Nelder e Wedderburn, 1972) estabelece uma relação entre a média da variável dependente Y com uma ou mais variáveis independentes x_i . Tem como característica o fato de que a distribuição da variável dependente tem que obrigatoriamente pertencer à família exponencial.

Considere uma amostra $\{y_1, y_2, ..., y_n\}$ de uma distribuição na família exponencial. Então sua função de probabilidade ou função densidade de probabilidade pode ser descrita pela equação apresentada abaixo (Casella e Berger, 2010):

$$P(y_i|\boldsymbol{\eta}_i) = h(y_i)c(\boldsymbol{\eta}_i)exp\left\{\sum_{j=1}^{k} w_j(\boldsymbol{\eta}_i)t_j(y_i)\right\}. \tag{3.1}$$

As funções $h(y_i)$, $c(\boldsymbol{\eta}_i)$, $w_j(\boldsymbol{\eta}_i)$ $\forall j$ e $t_j(y_i)$ $\forall j$ são funções conhecidas. Algumas distribuições conhecidas que pertecem à família exponencial são: Bernoulli, binomial, Poisson, exponencial, gama e normal.

A modelagem em questão pode ser estruturada em três componentes: o aleatório (formado pelas variáveis aleatórias independentes), o sistemático (modelo proposto composto pelas variáveis preditoras lineares nos parâmetros) e a função de ligação que estabelece a ligação entre os componentes citados acima. Para um maior aprofundamento com relação aos modelos lineares generalizadas, ver Dobson (2002).

3.1.1 Regressão de Poisson

A distribuição de Poisson é muito utilizada para o caso em que a variável de interesse assume valores inteiros não negativos. Sua função de probabilidade é:

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, y = 0, 1, 2, 3, ... \tag{3.2}$$

Pode-se mostrar que sua esperança e variância são iguais ao seu respectivo parâmetro, ou seja, $E(Y|\mu) = V(Y|\mu) = \mu$:

$$E(Y|\mu) = \sum_{y=0}^{\infty}\left(y\frac{e^{-\mu}\mu^y}{y!}\right) = \sum_{y=1}^{\infty}\left[\frac{e^{-\mu}\mu^y}{(y-1)!}\right].$$

Fazendo $k = y - 1$ tem-se:

$$E(Y|\mu) = \sum_{k=0}^{\infty}\left(\frac{e^{-\mu}\mu^{k+1}}{k!}\right) = \mu\sum_{k=0}^{\infty}\left(\frac{e^{-\mu}\mu^k}{k!}\right) = \mu.$$

$$E(Y^2|\mu) = \sum_{y=0}^{\infty}\left(y^2\frac{e^{-\mu}\mu^y}{y!}\right) = \sum_{y=1}^{\infty}\left[y\frac{e^{-\mu}\mu^y}{(y-1)!}\right].$$

Novamente fazendo $k = y - 1$ tem-se:

$$\begin{aligned} E(Y^2|\mu) &= \sum_{k=0}^{\infty}\left[(k+1)\frac{e^{-\mu}\mu^{k+1}}{k!}\right] = \\ &= \mu\sum_{k=0}^{\infty}\left(k\frac{e^{-\mu}\mu^k}{k!}\right) + \mu\sum_{k=0}^{\infty}\left(\frac{e^{-\mu}\mu^k}{k!}\right) = \\ &= \mu E(Y|\mu) + \mu = \mu^2 + \mu \end{aligned}$$

$$V(Y|\mu) = E(Y^2|\mu) - [E(Y^2|\mu)]^2 = \mu^2 + \mu - \mu^2 = \mu$$

Conclui-se que $E(Y|\mu) = V(Y|\mu) = \mu$.

Tal distribuição pode ser derivada a partir de um conjunto de suposições que são chamadas de postulados de Poisson. O teorema que será apresentado abaixo foi retirado de Casella e Berger (2010) e ilustra as condições necessárias para que a variável aleatória Y_t seja uma distribuição de Poisson com parâmetro μt:

Teorema. *Para cada $t \geq 0$ considere Y_t uma variável aleatória assumindo valores inteiros com as seguintes propriedades:*

(1) $Y_0 = 0$,

(2) $s < t \Rightarrow Y_s$ *e* $Y_t - Y_s$ *são independentes* ,

(3) Y_s *e* $Y_{t+s} - Y_t$ *são indenticamente distribuídos*,

(4) $\lim_{t\to 0} \dfrac{P(Y_t = 1)}{t} = \mu$,

(5) $\lim_{t\to 0} \dfrac{P(Y_t > 1)}{t} = 0$.

Respeitando todas as condições apresentadas então para qualquer número inteiro positivo k,

$$P(Y_t = k|\mu) = \frac{e^{-\mu t}(\mu t)^k}{k!},$$

ou seja, $Y_t \sim$ Poisson(μt).

Considerando Y_t como o número de chegadas no período de 0 a t a condição (1) pode ser interpretada como iniciar o processo sem nenhuma chegada. A condição (2) implica que chegadas em períodos de tempo disjuntos são independentes, (3) implica que o número de chegadas depende somente do comprimento do período, nesse caso t, (4) implica que a probabilidade de chegada é proporcional ao comprimento do período caso o comprimento

seja pequeno e (5) implica que não há chegadas simultâneas. Maiores detalhes do processo podem ser encontrados em James (2008).

Em um modelo de regressão de Poisson, as variáveis dependentes são assumidas seguirem uma distribuição de Poisson cuja média, μ_i, está associada com variáveis explicativas por meio de uma função de ligação. Como dito anteriormente a distribuição de Poisson pertence a família exponencial uma vez que:

$$P(y_i|\mu_i) = \frac{1}{y_i!} exp\left\{-\mu_i\right\} exp\left\{y_i ln(\mu_i)\right\}, \text{ em que}$$

$$\begin{aligned} h(y_i) &= \frac{1}{y_i!}, \\ c(\mu_i) &= exp\left\{-\mu_i\right\}, \\ w_1(\mu_i) &= ln(\mu_i), \\ t_1(y_i) &= y_i. \end{aligned}$$

Como μ_i só pode assumir valores positivos é comum adotar uma função de ligação logarítmica, ou seja:

$$g(\mu_i) = ln(\mu_i) = \mathbf{X}_i\boldsymbol{\theta} = \theta_1 + \theta_2 x_{i1} + \theta_3 x_{i2} + ... + \theta_{n+1} x_{in}, \tag{3.3}$$

onde $\mathbf{X}$ é a matriz de desenho e θ_i são os coeficientes de regressão associados às variáveis explicativas. Equivalentemente, tem-se:

$$\mu_i = exp\left\{\mathbf{X}_i\boldsymbol{\theta}\right\} = exp\left\{\theta_1 + \theta_2 x_{i1} + \theta_3 x_{i2} + ... + \theta_{n+1} x_{in}\right\}. \tag{3.4}$$

Uma vez obtida uma amostra $y_i, i = 1, 2, ..., n$, o logaritmo da função de verossimilhança do modelo será:

$$ln[L(\boldsymbol{\theta})] = \sum_{i=1}^{n} \left[-exp\left\{\mathbf{X}_i\boldsymbol{\theta}\right\} + y_i\mathbf{X}_i\boldsymbol{\theta} - ln(y_i!)\right] \tag{3.5}$$

Usando a abordagem bayesiana determina-se uma distribuição *a priori* para $\boldsymbol{\theta}$. Assim o núcleo da distribuição *a posteriori* será:

$$P(\boldsymbol{\theta}|\mathbf{Y}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta}) \tag{3.6}$$

Essa expressão em geral não é analiticamente tratável. Então para obter uma amostra de $\boldsymbol{\theta}$ é necessário aplicar algum dos algoritmos apresentados anteriormente na seção 2.3.

3.2 Modelos lineares dinâmicos (MLD)

A classe dos modelos dinâmicos tem como característica permitir que os parâmetros evoluam ao longo do tempo, permitindo que se estime possíveis tendências e mesmo efeitos sazonais. Além de ter os componentes das séries diretamente interpretáveis, eles também conseguem indicar uma medida de incerteza associada às estimativas obtidas, além da capacidade adaptativa ao longo da amostra, através de um mecanismo de atualização de um período de tempo para o outro, gerando uma interpretação e estimativas para cada instante de tempo. Tais modelos são aplicados para dados normais. Foi feito um breve resumo da classe dos MLD. Toda parte teórica aqui mencionada pode ser encontrada em West e Harrison (1997).

O pressuposto básico dos modelos dinâmicos são que as observações vão flutuar em torno de uma média. Nos modelos estáticos essa média está fixa ao longo do tempo, mas em modelos dinâmicos tal média está sujeita a pequenas variações ao longo do tempo. Esse tipo de variação no sinal é essencialmente estocástico.

O processo de inferência, como mencionado anteriormente pode ser feito de maneira sequencial, ou seja, é recalculado a cada instante de tempo. Para estimar no instante $t = 1$ a informação utilizada está contida no conjunto D_0, que é o conjunto de informações disponíveis antes do tempo $t = 1$, podendo ser subjetivas ou não. Quando o processo chegar no instante $t = 2$ as informações utilizadas agora estão contidas no conjunto D_1, que pode ser interpretado como a união do conjunto D_0 com as novas informações obtidas. Desse modo tem-se $D_1 = \{D_0, I_1\}$, I_1 sendo o conjunto de informações obtidas

no instante $t = 1$. Esse processo é feito de maneira sucessiva obtendo assim estimativas para cada instante de tempo. Caso o conjunto D_t seja $D_t = \{D_0, y_1, y_2, ..., y_t\}$, ou seja, em cada passagem de tempo a única informação incluída for y_t, diz-se que o sistema é fechado.

Utilizando a abordagem bayesiana o modelo para dados normais pode ser formalizado pelos seguintes componentes:

$$\text{equação de observação} \quad : \quad \mathbf{y}_t = \mathbf{F}_t'\boldsymbol{\theta}_t + \mathbf{v}_t, \text{ onde } \mathbf{v}_t \sim NM\left(\mathbf{0}, \mathbf{V_t}\right), \tag{3.7}$$

$$\text{equação de evolução} \quad : \quad \boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \text{ onde } \mathbf{w}_t \sim NM\left(\mathbf{0}, \mathbf{W}_t\right), \tag{3.8}$$

$$\text{informação inicial} \quad : \quad \boldsymbol{\theta}_0|D_0 \sim NM\left(\mathbf{m}_0, \mathbf{C}_0\right). \tag{3.9}$$

Para $t = 1, 2, ..., T$, tem-se que $\mathbf{y}_t$ é o vetor de observações de dimensão $p \times 1$, $\mathbf{F}_t'$ é uma matriz conhecida de dimensão $p \times n$, $\boldsymbol{\theta}_t$ é o conjunto de parâmetros do modelo (denominados parâmetros de estados) de dimensão $n \times 1$, $\mathbf{V}_t$ é uma matriz de covariâncias conhecida de dimensão $p \times p$, $\mathbf{G}_t$ é uma matriz conhecida de dimensão $n \times n$, $\mathbf{W}_t$ é a matriz de covariâncias também conhecida de dimensão $n \times n$ e $\boldsymbol{\theta}_0|D_0$ é a distribuição normal multivariada *a priori* condicionada a informação inicial.

A evolução dos parâmetros é controlada através do termo aleatório $\mathbf{w}_t$. Note que quanto maior a variância de $\mathbf{w}_t$ maior será a variação dos valores dos parâmetros em instantes consecutivos de tempo. Em contrapartida a diminuição da variância faz com a variação dos valores dos parâmetros em instantes consecutivos de tempo fiquem muito pequena, tornando o modelo aproximadamente estático.

As distribuições $\mathbf{v}_t$ e $\mathbf{w}_t$ são assumidas independentes entre si para $t = 1, 2, ..., T$ e de $\mu_0|D_0$. Logo o modelo fica completamente definido pela quádrupla $\{\mathbf{F}_t, \mathbf{G}_t, \mathbf{V}_t, \mathbf{W}_t\}$. Um caso particular é quando $F_t' = 1$, $G_t = 1$ e $\theta_t = \mu_t$.

O processo de inferência sequencial é descrito pelo algoritmo conhecido como Filtro de Kalman. Tal algoritmo é descrito pelo conjunto de equações abaixo:

$$\mathbf{a}_t = E(\boldsymbol{\theta}_t|D_{t-1}) = \mathbf{G}_t\mathbf{m}_{t-1}, \tag{3.10}$$

$$\mathbf{R}_t = V(\boldsymbol{\theta}_t|D_{t-1}) = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t, \tag{3.11}$$

$$\mathbf{m}_t = E(\boldsymbol{\theta}_t|D_t) = \mathbf{a}_t + \frac{\mathbf{R}_t\mathbf{F}_t(\mathbf{y}_t - \mathbf{F}_t'\mathbf{a}_t)}{\mathbf{Q}_t}, \tag{3.12}$$

$$\mathbf{C}_t = V(\boldsymbol{\theta}_t|D_t) = \mathbf{R}_t + \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t}{\mathbf{Q}_t}. \tag{3.13}$$

onde $\mathbf{Q}_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + \mathbf{V}_t$.

As duas primeiras equações são responsáveis pela predição, obtendo assim as estimativas para o vetor de parâmetros $\boldsymbol{\theta}_t$ e sua matriz de covariância $\mathbf{W}_t$. Note que nessa etapa utilizou-se apenas as informações disponíveis até o instante $t-1$. As últimas equações são chamadas equações de atualização pois elas atualizam as estimativas obtidas utilizando o conjunto de dados D_t.

Uma característica importante do estimador gerado pelo Filtro de Kalman é que ele minimiza o erro quadrático médio de previsão dentre todos os estimadores lineares sendo que sob a hipótese dos resíduos serem normalmente distribuídos estente-se a propriedade para todos os estimadores.

O filtro de Kalman pode ser descrito em termos das distribuições *priori* e *posteriori* para o vetor paramétrico $\boldsymbol{\theta}_t$:

$$\begin{aligned}
\text{Distribuição } \textit{posteriori} \text{ em } t-1 &: \boldsymbol{\theta}_{t-1}|D_{t-1} \sim NM(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}), \\
\text{Distribuição } \textit{priori} \text{ em } t &: \boldsymbol{\theta}_t|D_{t-1} \sim NM(\mathbf{a}_t, \mathbf{R}_t), \\
\text{Distribuição } \textit{posteriori} \text{ em } t &: \boldsymbol{\theta}_t|D_t \sim NM(\mathbf{m}_t, \mathbf{C}_t).
\end{aligned}$$

Para obter a distribuição preditiva $\mathbf{y}_t|D_{t-1}$ basta resolver a seguinte integral:

$$\begin{aligned}
p(\mathbf{y}_t|D_{t-1}) &= \int P(\mathbf{y}_t, \boldsymbol{\theta}_t|D_{t-1})d\boldsymbol{\theta}_t = \\
&= \int P(\boldsymbol{\theta}_t|D_{t-1})P(\mathbf{y}_t|\boldsymbol{\theta}_t, D_{t-1})d\boldsymbol{\theta}_t.
\end{aligned}$$

Sob a hipótese de normalidade de $\boldsymbol{\theta}_t|D_{t-1}$ e $\mathbf{y}_t|\boldsymbol{\theta}_t, D_{t-1}$ pode-se resolver a integral de maneira analítica obtendo assim:

$$\mathbf{y}_t|D_{t-1} \sim NM\left(\mathbf{f}_t, \mathbf{Q}_t\right), \tag{3.14}$$

onde $\mathbf{f}_t = \mathbf{F}_t'\mathbf{a}_t$ e $\mathbf{Q}_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + \mathbf{V}_t$.

Para obter a distribuição preditiva k passos a frente a partir do instante t obtém-se primeiro a distribuição $\boldsymbol{\theta}_{t+k}|D_t$:

$$\boldsymbol{\theta}_{t+k}|D_t \sim NM\left[\mathbf{a}_t(k), \mathbf{R}_t(k)\right],$$

onde $\mathbf{a}_t(k) = \mathbf{G}_{t+k}\mathbf{a}_t(k-1)$ e $\mathbf{R}_t(k) = \mathbf{G}_{t+k}\mathbf{R}_t(k-1)\mathbf{G}_{t+k}' + \mathbf{W}_{t+k}$. Os valores iniciais $\mathbf{a}_t(0)$ e $\mathbf{R}_t(0)$ são $\mathbf{m}_t$ e $\mathbf{C}_t$ respectivamente. Logo a preditiva k passos a frente a partir do instante t é dada por:

$$\mathbf{y}_{t+k}|D_t \sim NM\left[\mathbf{f}_t(k), \mathbf{Q}_t(k)\right], \tag{3.15}$$

onde $\mathbf{f}_t(k) = \mathbf{F}_{t+k}'\mathbf{a}_t(k)$ e $\mathbf{Q}_t(k) = \mathbf{F}_{t+k}'\mathbf{R}_t(k)\mathbf{F}_{t+k} + \mathbf{V}_{t+k}$.

O ciclo de inferência e previsão pode ser expresso por:

$$\begin{array}{ccccc} \boldsymbol{\theta}_{t-1}|D_{t-1} & \overset{\text{Evolução}}{\Rightarrow} & \boldsymbol{\theta}_t|D_{t-1} & \overset{\text{Atualização}}{\Rightarrow} & \boldsymbol{\theta}_t|D_t \\ & & \Downarrow & & \\ & & \mathbf{y}_t|D_{t-1} & & \\ & & \text{Previsão} & & \end{array}$$

3.3 Modelos lineares dinâmicos generalizados (MLDG)

Na seção anterior foi feito um breve resumo sobre os modelos lineares dinâmicos. Tais modelos são aplicados para dados supostamente normais. Em muitas situações não é razoável utilizar a hipótese de normalidade, logo a aplicação dessa classe de modelos não é recomendável. Para contornar este problema pode-se utilizar uma transformação

nos dados para que a suposição de normalidade seja plausível. Outra opção, considerada por muitos preferível, é trabalhar com os dados na escala original.

Diferente da modelagem apresentada anteriromente, em modelos lineares dinâmicos generalizados (West et al., 1985) a distribuição da variável de interesse é assumida pertencer à família exponencial. Desse modo tem-se que os MLDG são uma classe de modelos mais abrangentes podendo inclusive ser aplicados para dados discretos, contanto que as distribuições pertençam à família exponencial. Nesse seção foi feito um resumo dos MLDG. Toda parte teórica aqui mencionada pode ser encontrada em West e Harrison (1997).

A descrição do modelo pode ser formalizada pelos seguintes componentes:

$$\text{f.p. ou f.d.p.} \quad : \quad P\left(y_t|\eta_t, V_t\right) = b\left(y_t, V_t\right) exp\left\{\phi_t\left[Y_t\left(y_t\right)\eta_t - a\left(\eta_t\right)\right]\right\} \tag{3.16}$$

$$\text{equação de ligação} \quad : \quad g\left(\eta_t\right) = \lambda_t = \mathbf{F}_t'\boldsymbol{\theta}_t; \tag{3.17}$$

$$\text{equação de evolução} \quad : \quad \boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \text{ onde } \mathbf{w}_t \sim NM\left(\mathbf{0}, \mathbf{W}_t\right); \tag{3.18}$$

$$\text{informação inicial} \quad : \quad \boldsymbol{\theta}_0|D_0 \sim NM\left(\mathbf{m}_0, \mathbf{C}_0\right). \tag{3.19}$$

Para $t = 1, 2, ..., T$ tem-se que $\boldsymbol{\theta}_t$ é o conjunto de parâmetros do modelo de dimensão $n \times 1$, $\mathbf{F}_t'$ é uma matriz conhecida de dimensão $p \times n$, $\mathbf{G}_t$ é uma matriz conhecida de dimensão $n \times n$, $\mathbf{W}_t$ é a matriz de covariâncias também conhecida de dimensão $n \times n$, λ_t é uma função linear do vetor $\boldsymbol{\theta}_t$ e $g\left(\eta_t\right)$ uma função monótona contínua conhecida.

Condicionalmente a V_t, assume-se uma distribuição *a priori* $P\left(\eta_t|V_t, D_{t-1}\right)$ para η_t. Para fins de notação, denotou-se $P\left(\eta_t|V_t, D_{t-1}\right)$ por $P\left(\eta_t|D_{t-1}\right)$.

As distribuições $\mathbf{w}_t$ são assumidas normais independentes de y_t para $t = 1, 2, ..., T$ condicionais a η_t.

O processo de inferência dos MLDG é simililar ao processo apresentado na seção anterior. A diferença é que agora nem sempre a distribuição *a posteriori* será analiticamente tratável. Por esse motivo agora as distribuições serão parcialmente especificadas por suas médias e variâncias:

$$\text{Distribuição } \textit{posteriori} \text{ em } t-1 \quad : \quad \boldsymbol{\theta}_{t-1}|D_{t-1} \sim \left[\mathbf{m}_{t-1}, \mathbf{C}_{t-1}\right], \tag{3.20}$$

$$\text{Distribuição } \textit{priori} \text{ em } t \quad : \quad \boldsymbol{\theta}_t|D_{t-1} \sim \left[\mathbf{a}_t, \mathbf{R}_t\right], \tag{3.21}$$

onde $\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$ e $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t$.

Como mencionado anteriormente especifica-se a distribuição *a priori* $\eta_t|D_{t-1}$. Como $g(\eta_t) = \lambda_t = \mathbf{F}_t'\boldsymbol{\theta}_t$ pode-se obter também a distribuição $\lambda_t|D_{t-1}$ ou ainda a distribuição conjunta $\lambda_t, \boldsymbol{\theta}_t|D_{t-1}$ especificada pelo vetor de médias e matriz de covariância:

$$\lambda_t, \boldsymbol{\theta}_t|D_{t-1} \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} q_t & \mathbf{F}_t'\mathbf{R}_t \\ \mathbf{R}_t\mathbf{F}_t & \mathbf{R}_t \end{pmatrix} \right],$$

em que $f_t = \mathbf{F}_t'\mathbf{a}_t$ e $q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t$.

Uma vez observado o valor y_t, a distribuição de interesse é $\lambda_t|D_t$, que pode ser obtida pela atualização do modelo dada por:

$$\lambda_t|D_t \sim [f_t^*, q_t^*],$$

onde $f_t^* = f_t + (y_t - f_t)\frac{q_t}{q_t+V_t}$ e $q_t^* = q_t - \frac{q_t^2}{q_t+V_t}$.

O núcleo da distribuição de $\boldsymbol{\theta}_t|D_t$ pode ser obtida via teorema de Bayes. Tal núcleo é obtido a partir do núcleo da distribuição conjunta de $\lambda_t, \boldsymbol{\theta}_t|D_t$:

$$\begin{aligned} P(\lambda_t, \boldsymbol{\theta}_t|D_t) &\propto P(\lambda_t, \boldsymbol{\theta}_t|D_{t-1})P(y_t|\lambda_t) \\ &\propto P(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})P(\lambda_t, |D_{t-1})P(y_t|\lambda_t) \\ &\propto P(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})P(\lambda_t, |D_t) \end{aligned}$$

Dado η_t, $\boldsymbol{\theta}_t$ é condicionalmente independente de y_t. Logo obtém-se a distribuição *a posteriori* para $\boldsymbol{\theta}|D_t$:

$$\begin{aligned} P(\boldsymbol{\theta}_t|D_t) &= \int P\left(\lambda_t, \boldsymbol{\theta}_t|D_{t-1}\right) d\lambda_t \\ &= \int P\left(\boldsymbol{\theta}_t|\lambda_t, D_{t-1}\right) P\left(\lambda_t, |D_t\right) d\lambda_t \end{aligned}$$

O primeiro termo da integral pode ser parcialmente definido por sua média e variância. O cálculo em questão não é obtido de forma analítica. Então estima-se a média e variância utilizando o estimador linear de Bayes. Os valores ótimos são:

$$\begin{aligned}\hat{E}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1}) &= \mathbf{a}_t + \frac{\mathbf{R}_t\mathbf{F}_t(\lambda_t - f_t)}{q_t},\\ \hat{V}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1}) &= \mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t}{q_t}.\end{aligned}$$

O segundo termo da integral foi parcialmente especificado anteriormente.

Finalmente pode-se especificar parcialmente a distribuição $P(\boldsymbol{\theta}_t|D_t)$:

$$\begin{aligned}E(\boldsymbol{\theta}_t|D_t) &= E[E(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t],\\ V(\boldsymbol{\theta}_t|D_t) &= E[V(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t] + V[E(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t].\end{aligned}$$

Assim:

$$\boldsymbol{\theta}_t|D_t \sim [\mathbf{m}_t, \mathbf{C}_t]\,, \mathbf{m}_t = E[\hat{E}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t] = \mathbf{a}_t + \frac{\mathbf{R}_t\mathbf{F}_t(f_t^* - f_t)}{q_t} \text{ e}$$

$$\mathbf{C}_t = E[\hat{V}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t] + V[\hat{E}(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})|D_t] = \mathbf{R}_t - \frac{\mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t\left(1 - \frac{q_t^*}{q_t}\right)}{q_t}.$$

Para a obtenção da distribuição preditiva a um passo a frente deve-se tomar algumas medidas. A primeira é assumir que $\lambda_t|D_{t-1}$ é aproximadamente normal. A segunda é trabalhar com *prioris* conjugadas aos valores especificados para a distribução *a priori* de λ_t. Nesse caso a distribuição *a priori* tem a forma:

$$P(\eta_t|D_{t-1}) = c(r_t, s_t)exp\left\{r_t\eta_t - s_t a(\eta_t)\right\} \tag{3.22}$$

Desse modo a distribuição preditiva a um passo a frente será:

$$P(y_t|D_{t-1}) = \frac{c(r_t, s_t)b(y_t, V_t)}{c(r_t + \phi_t y_t, s_t + \phi_t)} \tag{3.23}$$

De modo análogo a distribuição a k passos a frente será:

$$P(y_{t+k}|D_t) = \frac{c(r_t(k), s_t(k))b(y_{t+k}, V_{t+k})}{c(r_t(k) + \phi_{t+k}y_{t+k}, s_t(k) + \phi_{t+k})} \tag{3.24}$$

3.3.1 Modelo Poisson dinâmico

Suponha que $Y_1, Y_2, ..., Y_t$ sejam variáveis aleatórias condicionalmente independentes onde $Y_t|\mu_t \sim$ Poisson(μ_t), para $t = 1, 2, ..., T$. A descrição do modelo pode ser formalizada por:

$$\begin{aligned}
\text{f.p.} &: P\left(y_t|\eta_t, V_t\right) = \frac{1}{y_t!} exp\left\{y_t ln(\mu_t) - \mu_t\right\} \\
\text{equação de ligação} &: g\left(\eta_t\right) = ln(\mu_t) = \mathbf{F}_t'\boldsymbol{\theta}_t; \\
\text{equação de evolução} &: \boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \text{ onde } \mathbf{w}_t \sim NM\left(\mathbf{0}, \mathbf{W}_t\right); \\
\text{informação inicial} &: \boldsymbol{\theta}_0|D_0 \sim NM\left(\mathbf{m}_0, \mathbf{C}_0\right).
\end{aligned}$$

Nesse caso a função de ligação é $ln(.)$, onde $\eta_t = \mu_t$; $\boldsymbol{\theta}_t$ é o vetor dos parâmetros de estado; $\mathbf{F}_t'$ e $\mathbf{G}_t$ são matrizes conhecidas e assume-se $\mathbf{W}_t = \mathbf{W}$ para $t = 1, 2, ..., T$. Assumiu-se distribuição *a priori* $\boldsymbol{\theta}_0|D_0 \sim NM\left(\mathbf{m}_0, \mathbf{C}_0\right)$, onde $\mathbf{m}_0$ e $\mathbf{C}_0$ são conhecidos e refletem a incerteza a respeito do processo no instante inicial.

Em muitas ocasiões o valor de $\mathbf{W}$ não é conhecido sendo necessário estimá-lo. Nesses situações atribui-se uma distribuição *a priori* para $\mathbf{W}$ dado a informação inicial D_0, ou seja, $\mathbf{W}|D_0$.

Capítulo 4

Modelos para placares de partidas de futebol

4.1 Estrutura geral dos modelos

Assim como nos modelos propostos por Dixon e Coles (1997); Souza Júnior e Gamerman (2004); Farias (2008); Gardner (2011), para modelar o placar de uma partida de futebol onde o time i enfrenta o time j, os números de gols de cada equipe são assumidos serem condicionalmente independentes cada um com distribuição de Poisson de forma que $Y_i^t|\lambda_i^t \sim Poisson(\lambda_i^t)$ e $Y_j^t|\lambda_j^t \sim Poisson(\lambda_j^t)$, em que Y_i^t, com média λ_i^t, é o número de gols do time i jogando como mandante na rodada t e Y_j^t, com média λ_j^t, é o número de gols do time j como visitante na rodada t para i, $j \in \{1, 2, ..., m\}$ e $t = 1, ..., T$, em que m é o número de times e T é o número de rodadas.

As médias, por sua, vez são assumidas compostas por três fatores: a força de ataque (α_i^{*t}), a força de defesa (β_i^{*t}) e o fator quando a equipe joga em casa (γ_i^{*t}). Os fatores se relacionam com a média do número de gols dos times mandante e visitante , respectivamente, por meio de funções de ligação da forma (Souza Júnior e Gamerman, 2004):

$$log(\lambda_i^t) = \alpha_i^{*t} - \beta_j^{*t} + \gamma_i^{*t}, \quad (4.1)$$

$$log(\lambda_j^t) = \alpha_j^{*t} - \beta_i^{*t}, \quad (4.2)$$

Note que uma vez determinada a rodada t e o time mandante i, o correspondente time adversário visitante j está determinado, assim como a rodada t e o time visitante j determinam o correspondente time adversário mandante i.

Para que seja possível estimar os fatores do modelo sem que haja problemas de identificabilidade, serão consideradas duas parametrizações. A primeira foi proposta por Farias (2008). Agora as equações de ligação contam com um nível comum a todos os times na rodada t, μ^t:

$$log(\lambda_i^t) = \mu^t + \alpha_i^t - \beta_j^t + \gamma_i^t, \quad (4.3)$$

$$log(\lambda_j^t) = \mu^t + \alpha_j^t - \beta_i^t, \quad (4.4)$$

onde $\mu^t = \alpha_1^{*t} - \beta_1^{*t}$, $\alpha_i^t = \alpha_i^{*t} - \alpha_1^{*t}$, $\beta_i^t = \beta_i^{*t} - \beta_1^{*t}$ e $\gamma_i^t = \gamma_i^{*t}$.

A segunda foi proposta por Gardner (2011). Nela considera-se α_1^{*t} como o fator base do modelo deixando assim de ser estimado. Todos os outros fatores de ataque e defesa estimados são comparados a esse fator base, ou seja, mede-se a diferença das forças desses fatores:

$$log(\lambda_i^t) = \alpha_i^t - \beta_j^t + \gamma_i^t, \quad (4.5)$$

$$log(\lambda_j^t) = \alpha_j^t - \beta_i^t, \quad (4.6)$$

onde $\alpha_i^t = \alpha_i^{*t} - \alpha_1^{*t}$, $\beta_i^t = \beta_i^{*t} - \alpha_1^{*t}$ e $\gamma_i^t = \gamma_i^{*t}$. Caso alguma estimativa dos coeficientes de ataque e defesa obtida seja próxima de zero significa que o coeficiente não difere do coeficiente base. Note que para essa parametrização também deixou-se de estimar um fator. Teoricamente, as estimativas para os fatores considerando as diferentes parametrizações não alteram as estimativas do logaritmo das médias. O que difere é o modo como elas são obtidas. Foi considerada nos modelos que serão apresentados nessa seção

a parametrização proposta por Gardner (2011), uma vez que nela comparou-se os fatores de ataque e defesa a apenas um fator.

Como dito anteriormente o procedimento de inferência será feito sob o enfoque bayesiano. As distribuições *a posteriori* dos modelos apresentados neste trabalho não possuem forma analítica fechada. Sendo assim, utilizou-se o método de Monte Carlo via cadeias de Markov (MCMC), em particular utilizando os algoritmos amostrador de Gibbs e Metropolis-Hastings. Para cada seção a seguir, apresentou-se um modelo proposto para a modelagem dos placares do campeonato brasileiro 2017, assim como detalhes do procedimento de inferência.

4.2 Modelo estático (ME)

Modelo aplicado por Souza Júnior e Gamerman (2004) em que os fatores de ataque, defesa e casa são assumidos serem estáticos ao longo das rodadas. O vetor transposto de parâmetros dos m times é:

$$\boldsymbol{\Theta}^{\intercal} = (\alpha_2, \alpha_3, ..., \alpha_m, \beta_1, \beta_2, ..., \beta_m, \gamma_1, \gamma_2, ..., \gamma_m)\,, \tag{4.7}$$

onde m é o número de times participantes do campeonato.

Atribui-se as seguintes distribuições *a priori* para os fatores de ataque, defesa e casa:

$$\begin{aligned}
\alpha_i &\sim Normal\left(\varphi_{\alpha_i}, \epsilon_{\alpha_i}\right), \text{ para } i = 2, 3, ..., m; \\
\beta_i &\sim Normal\left(\varphi_{\beta_i}, \epsilon_{\beta_i}\right) \text{ para } i = 1, 2, ..., m; \\
\gamma_i &\sim Normal\left(\varphi_{\gamma_i}, \epsilon_{\gamma_i}\right), \text{ para } i = 1, 2, ..., m;
\end{aligned}$$

em que $\varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ são constantes conhecidas para $\forall i$. Admitindo a independência *a priori* entre os fatores, tem-se a função de densidade conjunta dada por:

$$P(\boldsymbol{\Theta}) = \prod_{i=2}^{m} P(\alpha_i) \prod_{i=1}^{m} \left[P(\beta_i)P(\gamma_i)\right], \tag{4.8}$$

onde $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, ..., \alpha_m)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_m)$ e $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_m)$.

A função de verossimilhança é obtida pelo produto de todas as distribuições de número de gols de todas as partidas realizadas:

$$L(\boldsymbol{\Theta};\mathbf{Y}) = \prod_{t=1}^{T}\prod_{i=1}^{m} P\left(y_i^t|\lambda_i^t\right) = \prod_{t=1}^{T}\prod_{i=1}^{m}\left[\frac{\mathsf{e}^{-\lambda_i^t}\left(\lambda_i^t\right)^{y_i^t}}{y_i^t!}\right]. \tag{4.9}$$

Aplicando o teorema de Bayes, pode-se obter o núcleo da distribuição *a posteriori*, que é dado pelo produto da função de densidade *a priori* e a função de verossimilhança. Assim:

$$P(\boldsymbol{\Theta}|\mathbf{Y}) \propto P(\boldsymbol{\Theta})L(\boldsymbol{\Theta};\mathbf{Y}). \tag{4.10}$$

4.3 Modelo dinâmico (MD)

No modelo apresentado anteriormente, os fatores são estáticos, isto é, os parâmetros não variam no tempo. Em modelos dinâmicos permite-se que os fatores de ataque, defesa e casa variem dinamicamente no tempo, ou seja, ao longo das rodadas realizadas. Desse modo tem-se o vetor transposto de parâmetros da rodada t:

$$\left(\boldsymbol{\theta}^t\right)^{\intercal} = \left(\alpha_2^t, \alpha_3^t, ..., \alpha_m^t, \beta_1^t, \beta_2^t, ..., \beta_m^t, \gamma_1^t, \gamma_2^t, ..., \gamma_m^t\right). \tag{4.11}$$

.

Alguns exemplos de modelos dinâmicos para previsão de resultados que serão considerados são propostos por: Souza Júnior e Gamerman (2004) e Farias (2008). Para o modelo proposto por Souza Júnior e Gamerman (2004), assim como o proposto por Knorr-Held (2000), fatores de ataque, defesa e casa da equipe evoluem no tempo de acordo com as equações de evolução:

$$\begin{aligned}
\alpha_i^t &\sim Normal(\alpha_i^{t-1}, \sigma_{\alpha_i}^2),\\
\beta_i^t &\sim Normal(\beta_i^{t-1}, \sigma_{\beta_i}^2),\\
\gamma_i^t &\sim Normal(\gamma_i^{t-1}, \sigma_{\gamma_i}^2).
\end{aligned}$$

Para efeitos de simplificação do modelo assume-se $\sigma^2_{\alpha_i} = \sigma^2_{\beta_i} = \sigma^2_{\gamma_i} = \sigma^2 \; \forall i$, onde $W = \frac{1}{\sigma^2}$, em que $W \sim Gama(a, b)$ com a e b constantes conhecidas.

Assumindo que não existe informação antes da primeira rodada, as seguintes distribuições *a priori* para os parâmetros α_i^0, β_i^0 e γ_i^0 serão consideradas:

$$\begin{aligned}
\alpha_i^0 &\sim Normal\left(\varphi_{\alpha_i}, \epsilon_{\alpha_i}\right), \text{ para } i = 2, 3, ..., m; \\
\beta_i^0 &\sim Normal\left(\varphi_{\beta_i}, \epsilon_{\beta_i}\right), \text{ para } i = 1, 2, ..., m; \\
\gamma_i^0 &\sim Normal\left(\varphi_{\gamma_i}, \epsilon_{\gamma_i}\right), \text{ para } i = 1, 2, ..., m;
\end{aligned}$$

em que $\varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ são constantes conhecidas $\forall i$.

A distribuição *a priori* conjunta para $\boldsymbol{\theta} = \{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^T\}$ e W é dada por:

$$P(\boldsymbol{\theta}, W) = \prod_{t=1}^{T} \left[P(\boldsymbol{\theta}^t | \boldsymbol{\theta}^{t-1}, W)\right] P(\boldsymbol{\theta}^0) p(W), \tag{4.12}$$

onde $\boldsymbol{\theta}^t | \boldsymbol{\theta}^{t-1}, W \sim NM\left(\boldsymbol{\theta}^{t-1}, \mathbf{W}\right)$, em que $\mathbf{W} = \frac{1}{\sigma^2}\mathbf{I}$.

A função de verossimilhança do modelo dinâmico é parecida com a apresentada no modelo estático, com a diferença que agora ela está também em função do hiperparâmetro W.

$$L(\boldsymbol{\theta}, W; \mathbf{Y}) = \prod_{t=1}^{T} \prod_{i=1}^{m} P\left(y_i^t | \lambda_i^t\right) = \prod_{t=1}^{T} \prod_{i=1}^{m} \left[\frac{\mathrm{e}^{-\lambda_i^t} (\lambda_i^t)^{y_i^t}}{y_i^t!}\right]. \tag{4.13}$$

Aplicando o teorema de Bayes, tem-se:

$$P(\boldsymbol{\theta}, W | \mathbf{Y}) \propto P(\boldsymbol{\theta}, W) L(\boldsymbol{\theta}, W; \mathbf{Y}). \tag{4.14}$$

4.4 Modelo dinâmico com coeficientes auto-regressivos de evolução (MD1)

Modelo proposto por Farias (2008). Os fatores de ataque, defesa e casa da equipe evoluem no tempo de acordo com as equações de evolução:

$$\begin{aligned}
\alpha_i^t &\sim Normal(\phi_\alpha^t \alpha_i^{t-1}, \sigma_{\alpha_i}^2), \\
\beta_i^t &\sim Normal(\phi_\beta^t \beta_i^{t-1}, \sigma_{\beta_i}^2), \\
\gamma_i^t &\sim Normal(\phi_\gamma^t \gamma_i^{t-1}, \sigma_{\gamma_i}^2).
\end{aligned}$$

Novamente assume-se $\sigma_{\alpha_i}^2 = \sigma_{\beta_i}^2 = \sigma_{\gamma_i}^2 = \sigma^2 \ \forall i$, onde $W = \frac{1}{\sigma^2}$, em que $W \sim Gama(a, b)$ com a e b constantes conhecidas, $\phi_\alpha^t \sim Uniforme\,(0,1)$, $\phi_\beta^t \sim Uniforme\,(0,1)$ e $\phi_\gamma^t \sim Uniforme\,(0,1)$, para $t = 1, 2, ..., T$. A diferença com o modelo anterior é que o modelo em questão considera um coeficiente auto-regressivo para todos os fatores de ataque, um para os fatores de defesa e outro para os fatores casa em cada rodada t do campeonato.

Novamente será considerado que não existe informação antes da primeira rodada. Serão consideradas as seguintes distribuições *a priori* para os parâmetros α_i^0,β_i^0 e γ_i^0:

$$\begin{aligned}
\alpha_i^0 &\sim Normal\,(\varphi_{\alpha_i}, \epsilon_{\alpha_i})\,, \text{ para } i = 2, 3, ..., m; \\
\beta_i^0 &\sim Normal\,(\varphi_{\beta_i}, \epsilon_{\beta_i})\,, \text{ para } i = 1, 2, ..., m; \\
\gamma_i^0 &\sim Normal\,(\varphi_{\gamma_i}, \epsilon_{\gamma_i})\,, \text{ para } i = 1, 2, ..., m;
\end{aligned}$$

em que $\varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ são constantes conhecidas $\forall i$.

A distribuição *a priori* conjunta para $\boldsymbol{\theta} = \{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^T\}$, $\boldsymbol{\psi} = \{\boldsymbol{\phi}_\alpha, \boldsymbol{\phi}_\beta, \boldsymbol{\phi}_\gamma\}=$
$= \{\phi_\alpha^1, \phi_\alpha^2, ..., \phi_\alpha^T, \phi_\beta^1, \phi_\beta^2, ..., \phi_\beta^T, \phi_\gamma^1, \phi_\gamma^2, ..., \phi_\gamma^T\}$ e $W = \frac{1}{\sigma^2}$ é dada por:

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}, W) = \prod_{t=1}^{T} \left[P(\boldsymbol{\theta}^t | \boldsymbol{\theta}^{t-1}, \boldsymbol{\psi}^t, W) p(\boldsymbol{\psi}^t)\right] P(W) P(\boldsymbol{\theta}^0), \tag{4.15}$$

$\boldsymbol{\theta}^t | \boldsymbol{\theta}^{t-1}, \boldsymbol{\psi}^t, W \sim NM\left(\boldsymbol{\phi}^t \boldsymbol{\theta}^{t-1}, \mathbf{W}\right)$, $\mathbf{W} = \frac{1}{\sigma^2}\mathbf{I}$ e $\boldsymbol{\psi}^t = \left\{\phi_\alpha^t, \phi_\beta^t, \phi_\gamma^t\right\}$.

A função de verossimilhança do modelo auto-regressivo é:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}, W; \mathbf{Y}) = \prod_{t=1}^{T} \prod_{i=1}^{m} P\left(y_i^t | \lambda_i^t\right) = \prod_{t=1}^{T} \prod_{i=1}^{m} \left[\frac{\mathrm{e}^{-\lambda_i^t} \left(\lambda_i^t\right)^{y_i^t}}{y_i^t!}\right]. \tag{4.16}$$

Aplicando o teorema de Bayes, tem-se:

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}, W | \mathbf{Y}) \propto P(\boldsymbol{\theta}, \boldsymbol{\psi}, W) L(\boldsymbol{\theta}, \boldsymbol{\psi}, W; \mathbf{Y}). \tag{4.17}$$

4.5 Modelo dinâmico com coeficientes auto-regressivos de evolução com duas defasagens de tempo (MD2)

Com o objetivo de entender melhor a dependência temporal entre os fatores ao longo das rodadas em uma escala maior, propôs-se nesta dissertação um modelo em que os fatores de ataque, defesa e casa são assumidos evoluírem no tempo de acordo com as equações de evolução:

$$\begin{aligned}
\alpha_i^t &\sim Normal(\phi_{1\alpha}^t \alpha_i^{t-1} + \phi_{2\alpha}^t \alpha_i^{t-2}, \sigma_{\alpha_i}^2), \\
\beta_i^t &\sim Normal(\phi_{1\beta}^t \beta_i^{t-1} + \phi_{2\beta}^t \beta_i^{t-2}, \sigma_{\beta_i}^2), \\
\gamma_i^t &\sim Normal(\phi_{1\gamma}^t \gamma_i^{t-1} + \phi_{2\gamma}^t \gamma_i^{t-2}, \sigma_{\gamma_i}^2).
\end{aligned}$$

Assume-se $\sigma_{\alpha_i}^2 = \sigma_{\beta_i}^2 = \sigma_{\gamma_i}^2 = \sigma^2 \ \forall i$, onde $W = \frac{1}{\sigma^2}$, em que $W \sim Gama(a, b)$ com a e b constantes conhecidas, $\phi_{k\alpha}^t \sim Uniforme\,(0,1)$, $\phi_{k\beta}^t \sim Uniforme\,(0,1)$ e $\phi_{k\gamma}^t \sim Uniforme\,(0,1)$, para $k = 1, 2$ e $t = 2, 3, ..., T+1$. Devido às características do modelo proposto, vale ressaltar que a primeira rodada agora é representada por t=2; a segunda rodada por t=3; e assim por diante. Diferente do modelo anterior o modelo em questão considera dois coeficientes auto-regressivos, um para cada defasagem de tempo para todos os fatores de ataque, defesa e casa em cada rodada t do campeonato.

Serão assumidas as seguintes distribuições *a priori* para os parâmetros α_i^j, β_i^j e γ_i^j:

$$\begin{aligned}
\alpha_i^j &\sim Normal\,(\varphi_{\alpha_i}, \epsilon_{\alpha_i})\,, \text{ para } i = 2, 3, ..., m \text{ e } j = 0, 1; \\
\beta_i^j &\sim Normal\,(\varphi_{\beta_i}, \epsilon_{\beta_i})\,, \text{ para } i = 1, 2, ..., m \text{ e } j = 0, 1; \\
\gamma_i^j &\sim Normal\,(\varphi_{\gamma_i}, \epsilon_{\gamma_i}) \text{ para } i = 1, 2, ..., m \text{ e } j = 0, 1;
\end{aligned}$$

em que $\varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ pode-se considerar constantes conhecidas $\forall i, j$.

A distribuição *a priori* conjunta para $\boldsymbol{\theta} = \{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^T\}$, $\boldsymbol{\psi} = \left\{\boldsymbol{\phi}_{1\alpha}, \boldsymbol{\phi}_{1\beta}, \boldsymbol{\phi}_{1\gamma}, \boldsymbol{\phi}_{2\alpha}, \boldsymbol{\phi}_{2\beta}, \boldsymbol{\phi}_{2\gamma}\right\} =$ $= \left\{\phi_{1\alpha}^1, \phi_{1\alpha}^2, ..., \phi_{1\alpha}^T, \phi_{1\beta}^1, \phi_{1\beta}^2, ..., \phi_{1\beta}^T, \phi_{1\gamma}^1, \phi_{1\gamma}^2, ..., \phi_{1\gamma}^T, \phi_{2\alpha}^1, \phi_{2\alpha}^2, ..., \phi_{2\alpha}^T, \phi_{2\beta}^1, \phi_{2\beta}^2, ..., \phi_{2\beta}^T, \phi_{2\gamma}^1, \phi_{2\gamma}^2, ..., \phi_{2\gamma}^T\right\}$ e $W = \frac{1}{\sigma^2}$ é dada por:

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}, W) = \prod_{t=2}^{T+1} \left[P(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1}, \boldsymbol{\psi}^t, W)P(\boldsymbol{\psi}^t)\right] P(W)P(\boldsymbol{\theta}^0)P(\boldsymbol{\theta}^1), \tag{4.18}$$

$\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1}, \boldsymbol{\psi}^t, W \sim NM\left(\boldsymbol{\phi}_1^t\boldsymbol{\theta}^{t-1} + \boldsymbol{\phi}_2^t\boldsymbol{\theta}^{t-2}, \mathbf{W}\right)$, em que $\mathbf{W} = \frac{1}{\sigma^2}\mathbf{I}$ e $\boldsymbol{\psi}^t = \left\{\phi_{1\alpha}^t, \phi_{1\beta}^t, \phi_{1\gamma}^t, \phi_{2\alpha}^t, \phi_{2\beta}^t, \phi_{2\gamma}^t\right\}$.

A função de verossimilhança é:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}, W; \mathbf{Y}) = \prod_{t=2}^{T+1}\prod_{i=1}^{m} P\left(y_i^t|\lambda_i^t\right) = \prod_{t=2}^{T+1}\prod_{i=1}^{m}\left[\frac{\mathsf{e}^{-\lambda_i^t}\left(\lambda_i^t\right)^{y_i^t}}{y_i^t!}\right]. \tag{4.19}$$

Aplicando o teorema de Bayes, tem-se:

$$P(\boldsymbol{\theta}, \boldsymbol{\psi}, W|\mathbf{Y}) \propto P(\boldsymbol{\theta}, \boldsymbol{\psi}, W)L(\boldsymbol{\theta}, \boldsymbol{\psi}, W; \mathbf{Y}). \tag{4.20}$$

4.6 Modelo dinâmico com fatores estáticos e com coeficientes auto-regressivos de evolução (MDEST1)

O modelo com coeficientes auto-regressivos permite que estimar o quão dependentes no tempo são os fatores de ataque, defesa e casa. Entretanto, se há pouca dependência no tempo, esses fatores tendem a ficar em torno de zero. Considerou-se então uma variação dos modelo descrito na Seção 4.4 permitindo um nível comum a cada time. Desta forma, as equações de evolução dos fatores de ataque, defesa e casa são descritas como

$$\begin{aligned}
\alpha_i^t &\sim Normal(\alpha_i + \phi_\alpha^t\alpha_i^{t-1}, \sigma_{\alpha_i}^2),\\
\beta_i^t &\sim Normal(\beta_i + \phi_\beta^t\beta_i^{t-1}, \sigma_{\beta_i}^2),\\
\gamma_i^t &\sim Normal(\gamma_i + \phi_\gamma^t\gamma_i^{t-1}, \sigma_{\gamma_i}^2).
\end{aligned}$$

Assume-se $\sigma_{\alpha_i}^2 = \sigma_{\beta_i}^2 = \sigma_{\gamma_i}^2 = \sigma^2\ \forall i$, onde $W = \frac{1}{\sigma^2}$, em que $W \sim Gama(a, b)$ com a e b constantes conhecidas, $\phi_\alpha^t \sim Uniforme\,(0, 1)$, $\phi_\beta^t \sim Uniforme\,(0, 1)$ e $\phi_\gamma^t \sim Uniforme\,(0, 1)$, para $t = 1, 2, ..., T$. Note que sob a hipótese de independência entre os fatores no tempo, o modelo em questão se resume ao modelo estático, apresentado na Seção 4.2.

Serão consideradas as seguintes distribuições *a priori* para os parâmetros α_i, β_i, γ_i, α_i^0, β_i^0 e γ_i^0:

$$\begin{aligned}
\alpha_i &\sim Normal\left(\varphi_{\alpha_i}, \epsilon_{\alpha_i}\right), \text{ para } i = 2, 3, ..., m; \\
\beta_i &\sim Normal\left(\varphi_{\beta_i}, \epsilon_{\beta_i}\right) \text{ para } i = 1, 2, ..., m; \\
\gamma_i &\sim Normal\left(\varphi_{\gamma_i}, \epsilon_{\gamma_i}\right), \text{ para } i = 1, 2, ..., m; \\
\alpha_i^0 &\sim Normal\left(\varphi_{\alpha_i}, \epsilon_{\alpha_i}\right), \text{ para } i = 2, 3, ..., m; \\
\beta_i^0 &\sim Normal\left(\varphi_{\beta_i}, \epsilon_{\beta_i}\right), \text{ para } i = 1, 3, ..., m; \\
\gamma_i^0 &\sim Normal\left(\varphi_{\gamma_i}, \epsilon_{\gamma_i}\right), \text{ para } i = 1, 2, ..., m;
\end{aligned}$$

em que $\varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ são constantes conhecidas $\forall i$.

A distribuição *a priori* conjunta para $\boldsymbol{\theta}$, $\boldsymbol{\kappa} = (\alpha_2, \alpha_3, ..., \alpha_m, \beta_1, \beta_2, ..., \beta_m, \gamma_1, \gamma_2, ..., \gamma_m)$, $\boldsymbol{\psi}$ e $W = \frac{1}{\sigma^2}$ é dada por:

$$P(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W) = \prod_{t=1}^{T} \left[P(\boldsymbol{\theta}^t | \boldsymbol{\theta}^{t-1}, \boldsymbol{\kappa}, \boldsymbol{\psi}^t, W) p(\boldsymbol{\psi}^t)\right] \prod_{i=2}^{m} P(\alpha_i) \prod_{i=1}^{m} \left[P(\beta_i) P(\gamma_i)\right] P(W) P(\boldsymbol{\theta}^0), \tag{4.21}$$

$\boldsymbol{\theta}^t | \boldsymbol{\theta}^{t-1}, \boldsymbol{\kappa}, \boldsymbol{\psi}^t, W \sim NM\left(\boldsymbol{\kappa} + \boldsymbol{\phi}^t \boldsymbol{\theta}^{t-1}, \mathbf{W}\right)$, em que $\mathbf{W} = \frac{1}{\sigma^2}\mathbf{I}$ e $\boldsymbol{\psi}^t = \left\{\phi_\alpha^t, \phi_\beta^t, \phi_\gamma^t\right\}$.

A função de verossimilhança do é:

$$L(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W; \mathbf{Y}) = \prod_{t=1}^{T} \prod_{i=1}^{m} P\left(y_i^t | \lambda_i^t\right) = \prod_{t=1}^{T} \prod_{i=1}^{m} \left[\frac{\mathsf{e}^{-\lambda_i^t} \left(\lambda_i^t\right)^{y_i^t}}{y_i^t!}\right]. \tag{4.22}$$

Aplicando o teorema de Bayes, tem-se:

$$P(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W | \mathbf{Y}) \propto P(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W) L(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W; \mathbf{Y}). \tag{4.23}$$

4.7 Modelo dinâmico com fatores estáticos e com coeficientes auto-regressivos de evolução com duas defasagens de tempo (MDEST2)

Analogamente, uma variação do modelo com dois coeficientes auto-regressivos incluindo fatores estáticos também é proposta. Neste caso, os fatores de ataque, defesa e casa são assumidos evoluírem no tempo de acordo com as equações de evolução:

$$\begin{aligned}
\alpha_i^t &\sim Normal(\alpha_i + \phi_{1\alpha}^t \alpha_i^{t-1} + \phi_{2\alpha}^t \alpha_i^{t-2}, \sigma_{\alpha_i}^2), \\
\beta_i^t &\sim Normal(\beta_i + \phi_{1\beta}^t \beta_i^{t-1} + \phi_{2\beta}^t \beta_i^{t-2}, \sigma_{\beta_i}^2), \\
\gamma_i^t &\sim Normal(\gamma_i + \phi_{1\gamma}^t \gamma_i^{t-1} + \phi_{2\gamma}^t \gamma_i^{t-2}, \sigma_{\gamma_i}^2).
\end{aligned}$$

Assume-se $\sigma_{\alpha_i}^2 = \sigma_{\beta_i}^2 = \sigma_{\gamma_i}^2 = \sigma^2 \ \forall i$, onde $W = \frac{1}{\sigma^2}$, em que $W \sim Gama(a, b)$ com a e b constantes conhecidas, $\phi_{k\alpha}^t \sim Uniforme\,(0,1)$, $\phi_{k\beta}^t \sim Uniforme\,(0,1)$ e $\phi_{k\gamma}^t \sim Uniforme\,(0,1)$, para $k = 1, 2$ e $t = 2, 3, ..., T+1$.

Serão assumidas as seguintes distribuições *a priori* para os parâmetros α_i^j, β_i^j e γ_i^j:

$$\begin{aligned}
\alpha_i^j &\sim Normal\,(\varphi_{\alpha_i}, \epsilon_{\alpha_i})\,, \text{ para } i = 2, 3, ..., m \text{ e } j = 0, 1; \\
\beta_i^j &\sim Normal\,(\varphi_{\beta_i}, \epsilon_{\beta_i})\,, \text{ para } i = 1, 2, ..., m \text{ e } j = 0, 1; \\
\gamma_i^j &\sim Normal\,(\varphi_{\gamma_i}, \epsilon_{\gamma_i}) \text{ para } i = 1, 2, ..., m \text{ e } j = 0, 1;
\end{aligned}$$

em que $\varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ pode-se considerar constantes conhecidas $\forall i, j$. Além disso considera-se as seguintes distribuições a priori para α_i, β_i e γ_i:

$$\begin{aligned}
\alpha_i &\sim Normal\,(\varphi_{\alpha_i}, \epsilon_{\alpha_i})\,, \text{ para } i = 2, 3, ..., m; \\
\beta_i &\sim Normal\,(\varphi_{\beta_i}, \epsilon_{\beta_i}) \text{ para } i = 1, 2, ..., m; \\
\gamma_i &\sim Normal\,(\varphi_{\gamma_i}, \epsilon_{\gamma_i})\,, \text{ para } i = 1, 2, ..., m;
\end{aligned}$$

$\varphi_\mu = \varphi_{\alpha_i} = \varphi_{\beta_i} = \varphi_{\gamma_i} = \varphi$ e $\epsilon_{\alpha_i} = \epsilon_{\beta_i} = \epsilon_{\gamma_i} = \epsilon$ são constantes conhecidas para $\forall i$.

A distribuição *a priori* conjunta para $\boldsymbol{\theta}$, $\boldsymbol{\kappa}$, $\boldsymbol{\psi}$ e $W = \frac{1}{\sigma^2}$ é dada por:

$$P(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W) = \prod_{t=2}^{T+1} \left[P(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1}, \boldsymbol{\kappa}, \boldsymbol{\psi}^t, W)P(\boldsymbol{\psi}^t)\right] \prod_{i=2}^{m} P(\alpha_i) \prod_{i=1}^{m} \left[P(\beta_i)P(\gamma_i)\right] P(W)P(\boldsymbol{\theta}^0)P(\boldsymbol{\theta}^1), \tag{4.24}$$

$\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1}, \boldsymbol{\kappa}, \boldsymbol{\psi}^t, W \sim NM\left(\boldsymbol{\kappa} + \boldsymbol{\phi}_1^t\boldsymbol{\theta}^{t-1} + \boldsymbol{\phi}_2^t\boldsymbol{\theta}^{t-2}, \mathbf{W}\right)$,em que $\mathbf{W} = \frac{1}{\sigma^2}\mathbf{I}$, $\boldsymbol{\kappa} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$, $\boldsymbol{\psi}^t = \left\{\phi_{1\alpha}^t, \phi_{1\beta}^t, \phi_{1\gamma}^t, \phi_{2\alpha}^t, \phi_{2\beta}^t, \phi_{2\gamma}^t\right\}$.

A função de verossimilhança é:

$$L(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W; \mathbf{Y}) = \prod_{t=2}^{T+1} \prod_{i=1}^{m} P\left(y_i^t|\lambda_i^t\right) = \prod_{t=2}^{T+1} \prod_{i=1}^{m} \left[\frac{\mathsf{e}^{-\lambda_i^t}\left(\lambda_i^t\right)^{y_i^t}}{y_i^t!}\right]. \tag{4.25}$$

Aplicando o teorema de Bayes, tem-se:

$$P(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W|\mathbf{Y}) \propto P(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W)L(\boldsymbol{\theta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, W; \mathbf{Y}). \tag{4.26}$$

4.8 Modelo hierárquico estático (MHE)

Diferente dos modelos propostos nas seções anteriores, propôs-se um modelo em que o logaritmo do número de gols de um time é composto por variáveis relacionadas àquele time. Especificamente, as variáveis consideradas na modelagem são número de finalizações (X_{i1}^t), escanteios (X_{i2}^t), faltas (X_{i3}^t) e cartões (X_{i4}^t). A relação entre as médias do número de gols do time mandante (λ_i^t) e visitante (λ_j^t) são, respectivamente, dadas por:

$$log(\lambda_i^t) = \Phi_{i1}X_{i1}^t + \Phi_{i2}X_{i2}^t + \Phi_{i3}X_{i3}^t + \Phi_{i4}X_{i4}^t, \tag{4.27}$$

$$log(\lambda_j^t) = \Phi_{j1}X_{j1}^t + \Phi_{j2}X_{j2}^t + \Phi_{j3}X_{j3}^t + \Phi_{j4}X_{j4}^t. \tag{4.28}$$

As variáveis que se relacionam com as médias de gols, por sua vez, foram assumidas serem condicionalmente independentes seguindo uma distribuição de Poisson, isto é, $X_{ik}^t|\eta_{ik}^t \sim Poisson(\eta_{ik}^t)$ e $X_{jk}^t|\eta_{jk}^t \sim Poisson(\eta_{jk}^t)$. As médias, por sua, vez são assumidas compostas por três fatores: o fator (α_{ik}^*), o fator (β_{ik}^*) e o fator (γ_{ik}^*), análogos aos fatores

de ataque, defesa e casa. Novamente os fatores se relacionam com a média por meio de funções de ligação da forma:

$$log(\eta_{ik}^t) = \alpha_{ik} - \beta_{jk} + \gamma_{ik}, \tag{4.29}$$

$$log(\eta_{jk}^t) = \alpha_{jk} - \beta_{ik}. \tag{4.30}$$

onde $\alpha_{ik} = \alpha_{ik}^* - \alpha_{1k}^*$, $\beta_{ik} = \beta_{ik}^* - \alpha_{1k}^*$ e $\gamma_{ik} = \gamma_{ik}^*$. Observe que em cada parte referente as variáveis um fator α_{1k}^* foi utilizado como fator base para comparação aos demais fatores.

O vetor transposto dos parâmetros do modelo é:

$$\boldsymbol{\Theta}^\intercal = (\boldsymbol{\Phi}_{.1}, \boldsymbol{\Phi}_{.2}, \boldsymbol{\Phi}_{.3}, \boldsymbol{\Phi}_{.4}, \boldsymbol{\alpha}_{.1}, \boldsymbol{\alpha}_{.2}, \boldsymbol{\alpha}_{.3}, \boldsymbol{\alpha}_{.4}, \boldsymbol{\beta}_{.1}, \boldsymbol{\beta}_{.2}, \boldsymbol{\beta}_{.3}, \boldsymbol{\beta}_{.4}, \boldsymbol{\gamma}_{.1}, \boldsymbol{\gamma}_{.2}, \boldsymbol{\gamma}_{.3}, \boldsymbol{\gamma}_{.4}), \tag{4.31}$$

onde $\boldsymbol{\Phi}_{.k} = \{\Phi_{1k}, \Phi_{2k}, ..., \Phi_{mk}\}$, $\boldsymbol{\alpha}_{.k} = \{\alpha_{1k}, \alpha_{2k}, ..., \alpha_{mk}\}$, $\boldsymbol{\beta}_{.k} = \{\beta_{1k}, \beta_{2k}, ..., \beta_{mk}\}$ e $\boldsymbol{\gamma}_{.k} = \{\gamma_{1k}, \gamma_{2k}, ..., \gamma_{mk}\}$ para $k = 1, 2, 3, 4$.

Considerou-se a seguinte função de probabilidade conjunta para $\mathbf{Y}^t$, $\mathbf{X}_1^t$, $\mathbf{X}_2^t$, $\mathbf{X}_3^t$ e $\mathbf{X}_4^t$:

$$P(\mathbf{Y}^t; \mathbf{X}_1^t; \mathbf{X}_2^t; \mathbf{X}_3^t; \mathbf{X}_4^t | \boldsymbol{\Theta}) = \prod_{i=1}^{m} P\left(y_i^t | \lambda_i^t\right) \prod_{i=1}^{m} \prod_{k=1}^{4} P\left(x_{ik}^t | \eta_{ik}^t\right),$$

onde $\mathbf{X}_1^t = \{X_{11}^t, X_{21}^t, ..., X_{m1}^t\}$, $\mathbf{X}_2^t = \{X_{12}^t, X_{22}^t, ..., X_{m2}^t\}$, $\mathbf{X}_3^t = \{X_{13}^t, X_{23}^t, ..., X_{m3}^t\}$ e $\mathbf{X}_4^t = \{X_{14}^t, X_{24}^t, ..., X_{m4}^t\}$. Assim como no modelo proposto por Ma e Kockelman (2006), foi admitida independência condicional das variáveis número de finalizações, escanteios, faltas e cartões.

A função de verossimilhança do modelo é:

$$L(\boldsymbol{\Theta}; \mathbf{Y}; \mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3; \mathbf{X}_4) = \prod_{t=1}^{T} \prod_{i=1}^{m} P\left(y_i^t | \lambda_i^t\right) \prod_{t=1}^{T} \prod_{i=1}^{m} \prod_{k=1}^{4} P\left(x_{ik}^t | \eta_{ik}^t\right) =$$

$$= \prod_{t=1}^{T} \prod_{i=1}^{m} \left[\frac{\mathrm{e}^{-\lambda_i^t} \left(\lambda_i^t\right)^{y_i^t}}{y_i^t!} \right] \prod_{t=1}^{T} \prod_{i=1}^{m} \prod_{k=1}^{4} \left[\frac{\mathrm{e}^{-\eta_{ik}^t} \left(\eta_{ik}^t\right)^{x_{ik}^t}}{x_{ik}^t!} \right]. \tag{4.32}$$

Foram atribuídas as seguintes distribuições *a priori* para os parâmetros do modelo:

$$
\begin{aligned}
\Phi_{ik} &\sim Normal\left(\varphi_{\Phi_{ik}}, \epsilon_{\Phi_{ik}}\right), \text{ para } i = 1, 2, ..., m \text{ e } k = 1, 2, 3, 4; \\
\alpha_{ik} &\sim Normal\left(\varphi_{\alpha_{ik}}, \epsilon_{\alpha_{ik}}\right), \text{ para } i = 2, 3, ..., m \text{ e } k = 1, 2, 3, 4; \\
\beta_{ik} &\sim Normal\left(\varphi_{\beta_{ik}}, \epsilon_{\beta_{ik}}\right), \text{ para } i = 1, 2, ..., m \text{ e } k = 1, 2, 3, 4; \\
\gamma_{ik} &\sim Normal\left(\varphi_{\gamma_{ik}}, \epsilon_{\gamma_{ik}}\right), \text{ para } i = 1, 2, ..., m \text{ e } k = 1, 2, 3, 4;
\end{aligned}
$$

em que $\varphi_{\Phi_{ik}} = \varphi_{\alpha_{ik}} = \varphi_{\beta_{ik}} = \varphi_{\gamma_{ik}} = \varphi$ e $\epsilon_{\Phi_{ik}} = \epsilon_{\alpha_{ik}} = \epsilon_{\beta_{ik}} = \epsilon_{\gamma_{ik}} = \epsilon$ são constantes conhecidas $\forall i, k$. Admitindo independência entre as distribuições *prioris*, a distribuição *a priori* conjunta será:

$$
P(\boldsymbol{\Theta}) = \prod_{i=1}^{20}\prod_{k=1}^{4} P(\Phi_{ik}) \prod_{i=1}^{19}\prod_{k=1}^{4} [P(\alpha_{ik})] \prod_{i=1}^{20}\prod_{k=1}^{4} [P(\beta_{ik})P(\gamma_{ik})]
$$

Aplicando o teorema de Bayes, tem-se:

$$
P(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4) \propto P(\boldsymbol{\Theta}) L(\boldsymbol{\Theta}; \mathbf{Y}; \mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3; \mathbf{X}_4). \tag{4.33}
$$

Capítulo 5

Resultados

5.1 Introdução

Nesta seção, serão discutidos os resultados correspondentes à aplicação dos modelos apresentados no Capítulo anterior para dados do campeonato brasileiro. Especificamente, os dados correspondem a informações sobre placares e covariáveis do Campeonato Brasileiro da série A edição 2017 até a trigésima quinta rodada.

A parametrização considerada nos modelos foi a proposta por Gardner (2011). Como dito anteriormente nela considera-se α_1^{*t} como o fator base do modelo deixando assim de ser estimado. Todos os outros fatores de ataque e defesa estimados são comparados a esse fator base. Entretanto, em alguns modelos (MD, MD1 e MD2), para efeito de avaliação da convergência das cadeias, também foi aplicada a parametrização proposta por Farias (2008), com um nível μ_t comum para todas as equipes. Uma vez constatado que ambas obtiveram os mesmos resultados em relação a convergência optou-se aplicar a parametrização proposta por Gardner, uma vez que nela comparou-se os fatores de ataque e defesa a apenas um fator.

Na maioria dos modelos foi feita a previsão da rodada 36 do campeonato. Não foram feitas as previsões das rodadas 37 e 38, uma vez que essas duas últimas rodadas do campeonato apresentam complicações extras motivada pela definição de vários times na classificação, causando uma mudança de foco nos clubes. Destaca-se que o campeão já estava definido, algumas equipes já estavam classificadas para Libertadores como Cruzeiro-MG,

Grêmio-RS e Palmeiras-SP e o Flamengo-RJ estava disputando a final de outro torneio.

Importante destacar que nos modelos MD1, MD2, MDEST1 e MDEST2, apesar de todos pressuporem coeficientes auto-regressivos para cada rodada, para efeito de estimação e simplificação dos modelos, foram considerados únicos os coeficientes ao longo das rodadas, ou seja $\boldsymbol{\phi}^t = \boldsymbol{\phi}$, $\forall t$.

Para efeitos de notação adotou-se um índice numérico para os times do Campeonato Brasileiro edição 2017. A Tabela 5.1 descreve o índice escolhido, assim como sua sigla, dos times que será adotado nos modelos:

Tabela 5.1: Índices e siglas das equipes do Campeonato Brasileiro edição 2017

Equipe	Sigla	Índice	Equipe	Sigla	Índice
Atlético-GO	ACG	1	Flamengo-RJ	FLA	11
Atlético-MG	CAM	2	Fluminense-RJ	FLU	12
Atlético-PR	CAP	3	Grêmio-RS	GRE	13
Avaí-SC	AVA	4	Palmeiras-SP	PAL	14
Bahia-BA	BAH	5	Ponte Preta-SP	PON	15
Botafogo-RJ	BOT	6	Santos-SP	SAN	16
Chapecoense-SC	CHA	7	São Paulo-SP	SAO	17
Corinthians-SP	COR	8	Sport-PE	SPO	18
Coritiba-PR	CFC	9	Vasco da Gama-RJ	VAS	19
Cruzeiro-MG	CRU	10	Vitória-BA	VIT	20

Os números de gols são assumidos serem condicionalmente independentes cada um com distribuição de Poisson de forma que $Y_i^t|\lambda_i^t \sim Poisson\left(\lambda_i^t\right)$ e $Y_j^t|\lambda_j^t \sim Poisson\left(\lambda_j^t\right)$. As probabilidades de vitória, empate e derrota do time mandante na rodada t podem ser calculadas, respectivamente, a partir das equações seguintes:

$$P\left[\left(Y_i^t = k_i\right) > \left(Y_j^t = k_j\right)\right] = \sum_{k_i > k_j} \left[P\left(Y_i^t = k_i\right) P\left(Y_j^t = k_j\right)\right] \tag{5.1}$$

$$P\left[\left(Y_i^t = k_i\right) = \left(Y_j^t = k_j\right)\right] = \sum_{k_i = k_j} \left[P\left(Y_i^t = k_i\right) P\left(Y_j^t = k_j\right)\right] \tag{5.2}$$

$$P\left[\left(Y_i^t = k_i\right) < \left(Y_j^t = k_j\right)\right] = \sum_{k_i < k_j} \left[P\left(Y_i^t = k_i\right) P\left(Y_j^t = k_j\right)\right] \tag{5.3}$$

Como no campeonato não houve nenhum time que tenha realizado numa partida sete gols ou mais, optou-se utilizar as categorias 0, 1, 2, 3, 4, 5, 6 e 7 gols ou mais (7^+). Assim as probabilidades apresentadas anteriormente ficam como soma de parcelas finitas e a probabilidade $P\left[\left(Y_i^t = 7^+\right), \left(Y_j^t = 7^+\right)\right]$, que tende a zero em todos os casos, é incorporada na probabilidade de empate.

Como mencionado anteriormente utilizou-se o JAGS (Plummer, M, 2013), Just Another Gibbs Sampler, nessa dissertação. Aplicou-se o pacote rjags do software R Core Team (2017) para ajustar os modelos mencionados. Esse pacote permite a operação do JAGS através do R. Todas as análises estatísticas e os gráficos foram feitas também no software R.

Para todos os modelos foram obtidas duas cadeias, iniciando de valores diferentes, para cada parâmetro. Foram realizadas, exceto no modelo estático, 650000 iterações. Em seguida aplicou-se um *Burn-in* de 50000 iterações, ou seja, foram descartadas as primeiras 50000 interações de ambas as cadeias. Para eliminar a correlação entre as iterações restantes foi aplicado um espaçamento igual a 300, restando 2000 iterações para cada cadeia, totalizando ao final uma amostra de tamanho 4000. Vale destacar que no modelo estático foram realizadas 250000 iterações, com um *Burn-in* de 50000 e espaçamento de 100 iterações.

Foram admitidas as constantes $\varphi = 0$ e $\epsilon = \frac{1}{0.01}$ nos modelos ME e MHE referentes as distribuições *a priori* dos fatores dos times. Nos modelos restantes foram admitidas as mesmas constantes citadas anteriormente e $a = b = 0.1$, referentes a distribuição *a priori* da precisão.

Além dos parâmetros de interesse em cada modelo, obteve-se também amostras da distribuição a posteriori das probabilidades de vitória, empate e derrota. Para fins de

previsão, considerou-se a média *a posteriori* dessas probabilidades e aquela de maior valor foi adotada para indicar a previsão de resultado da rodada.

5.2 Análise descritiva

Todos os modelos apresentados anteriormente tem como objetivo estimar o número de gols realizados em uma partida de futebol. Para isso admite-se que o número de gols do mandante e do visitante seguem cada um uma distribuição de Poisson. A Figura 5.1 mostra a comparação das frequências relativas dos gols do campeonato (verde) com as probabilidades do número de gols obtidas pela Poisson (preto), cujos parâmetros λ são as médias obtidas das distribuições dos números de gols do mandante e visitante:

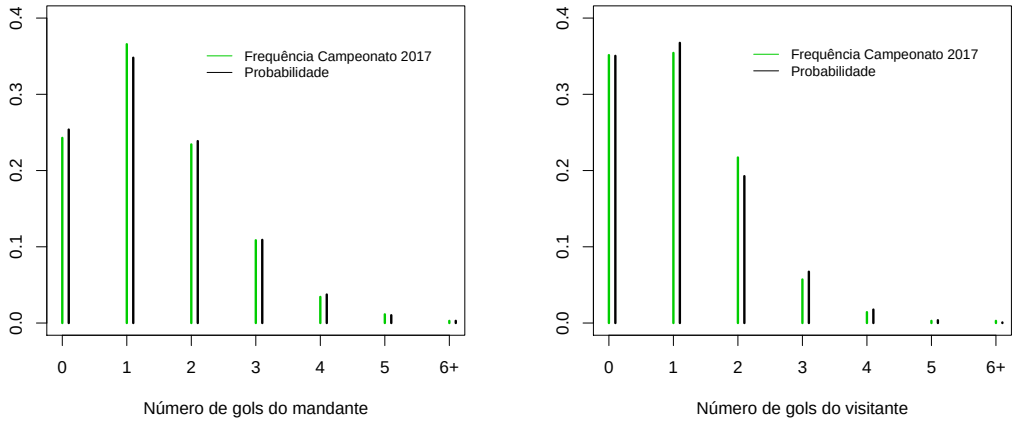


Figura 5.1: Comparação entre as distribuições dos números de gols dos times mandantes e visitantes com probabilidades obtidas das distribuições teóricas de Poisson.

Visualmente percebe-se que as frequências relativas estão muito próximas das probabilidades calculadas dos número de gols.

Apesar da abordagem do problema ser feita sob o enfoque bayesiano, lançou-se mão de alguns métodos clássicos de inferência para fins de análise exploratória dos dados. Para verificar se existe evidências estatísticas para rejeitar as hipóteses que as distribuições dos números de gols dos times mandantes e visitantes não são distribuições de

Poisson realizou-se o teste Qui-quadrado. Nele compara-se duas ou mais categorias independentes, cada uma com um respectivo tamanho. A hipótese nula (H_0) considera que a distribuição do número de gols segue uma distribuição de Poisson com parâmetro μ. Considerando um tamanho de amostra n suficientemente grande a estatística de teste é aproximadamente:

$$Q_{cal} = \sum_{i=1}^{k} \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}} \sim \chi^2_q,$$

em que f_{e_i} é a frequência esperada para a categoria i obtida da distribuição teórica, f_{o_i} é a frequência observada para a categoria i obtida da distribuição dos gols feitos e q representa os graus de liberdade. Uma apresentação mais rigorosa do teste pode ser encontrada em Murteira (1990).

As categorias determinadas para o teste foram 0, 1, 2, 3, 4, 5 e 6 ou mais gols, totalizando sete categorias. Nesse caso tem-se que os graus de liberdade serão $q = k-1-1 = 7-1-1 = 5$, uma vez que utilizou-se a estimativa de máxima verossimilhança do parâmetro μ para a obtenção da frequência esperada.

Os resultados obtidos para os gols do mandante e do visitante são apresentados na Tabela 5.2 abaixo:

Tabela 5.2: Teste Qui-quadrado

Distribuição	Estatística Teste (Q)	P-valor
Gols do mandante - 2017	0.6455	0.9858
Gols do visitante -2017	4.1218	0.5320

Em ambos os casos, ao nível de significância de 5%, não existem evidências estatísticas suficientes para rejeitar (H_0), ou seja, as hipóteses de que as distribuições do número de gols do mandante e visitante seguem distribuições de Poisson não podem ser rejeitadas.

5.3 Modelo estático (ME)

As figuras a seguir são os intervalos de 95% de credibilidade das amostras obtidas dos fatores de ataque, defesa e casa:

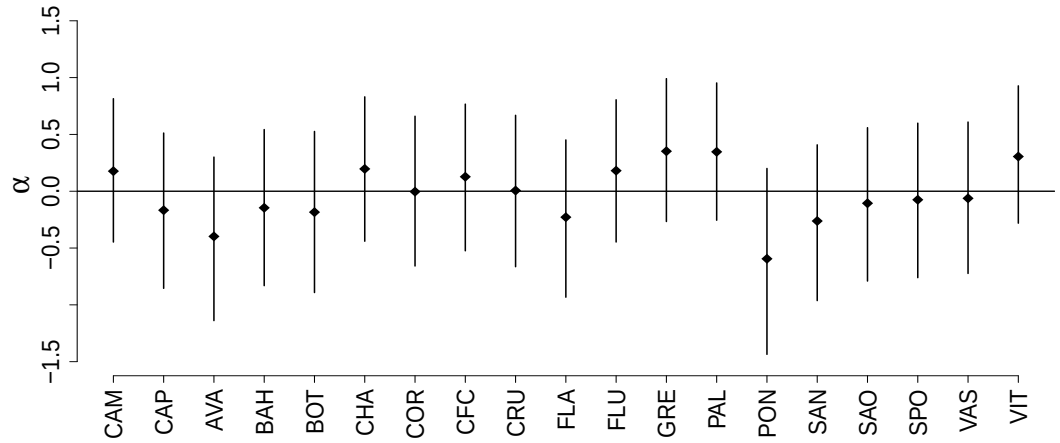


Figura 5.2: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de ataque do modelo ME.

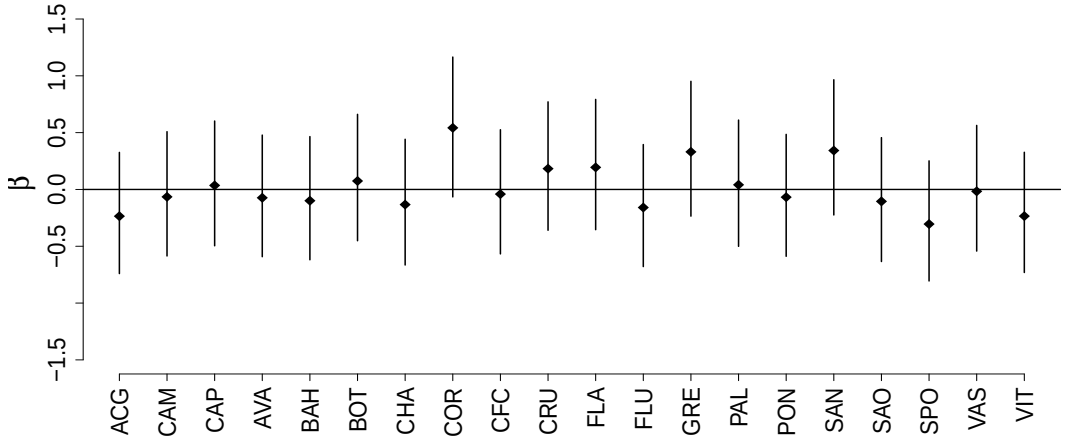


Figura 5.3: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de defesa do modelo ME.

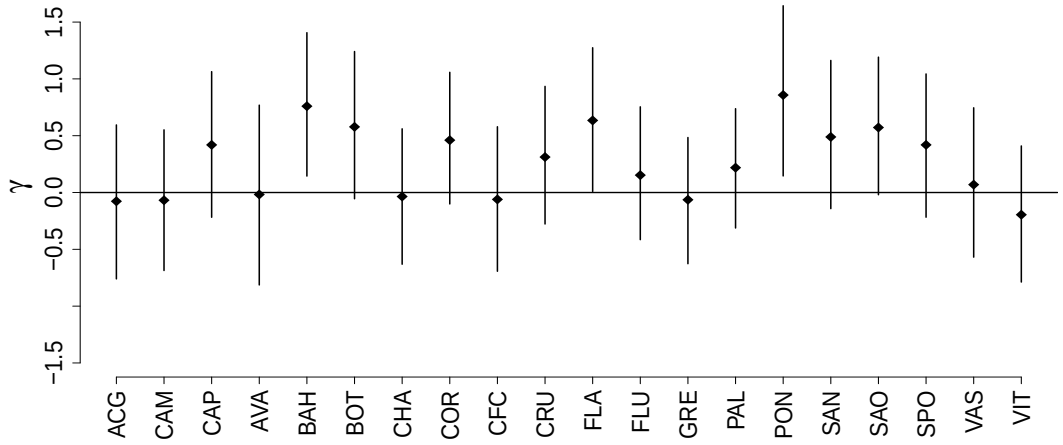


Figura 5.4: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores casa do modelo ME.

Com relação as médias referentes aos fatores de ataque ilustradas na Figura 5.2, destaca-se que os times que obtiveram melhores médias foram Grêmio-RS (0.3520), Palmeiras-SP (0.3467) e Vitória-BA (0.3052), enquanto Ponte Preta-SP (-0.5946), Avaí-SC (-0.3976) e Santos-SP (-0.2621) obtiveram as menores médias das equipes. Destaca-se da Figura 5.3 que o Corinthians-SP obteve melhor média (0.5420) dos fatores de defesa em relação as demais equipes seguido de Santos-SP (0.3424) e Grêmio-RS (0.3307). Sport-PE (-0.3046), Atlético-GO (-0.2354) e Vitória-BA (-0.2347) obtiveram as menores médias dos fatores de defesa. Por fim na Figura 5.4 nota-se que as equipes Ponte Preta-SP (0.8578), Bahia-BA (0.7596) e Flamengo-RJ (0.6344) obtiveram melhores médias referentes aos fatores casa e Vitória-BA (-0.1955), Atlético-GO (-0.0770) e Atlético-MG (-0.0681) obtiveram as menores médias.

O processo para obtenção das amostras dos números médios de gols das partidas pode ser descrito da seguinte forma: suponha que o time i jogará em casa contra o time j na rodada $t+1$. Obtém-se uma amostra, como descrito anteriormente, das distribuições *a posteriori* dos parâmetros usando os dados até a rodada t. Para cada elemento da amostra calcula-se λ_i^{t+1} e λ_j^{t+1}:

$$\lambda_i^{t+1} = exp\{\alpha_i - \beta_j + \gamma_i\} \quad (5.4)$$

$$\lambda_j^{t+1} = exp\{\alpha_j - \beta_i\} \quad (5.5)$$

A Tabela 5.3 ilustra as probabilidades médias obtidas da rodada 36 com seus respectivos intervalos de 95% de credibilidade:

Tabela 5.3: Médias *a posteriori* e respectivos intervalos de 95% de credibilidade *a posteriori* das probabilidades de vitória, empate e derrota para as partidas da 36ª rodada segundo o modelo ME

Rodada 36	Vitória	Empate	Derrota	Verificação do prognóstico
FLA 3X0 COR	$34.78\%_{[16.61\%;\,55.50\%]}$	$33.25\%_{[24.82\%;\,43.07\%]}$	$31.96\%_{[15.72\%;\,53.11\%]}$	Certo
SAO 0X0 BOT	$49.93\%_{[27.10\%;\,72.12\%]}$	$25.58\%_{[16.98\%;\,33.92\%]}$	$24.48\%_{[9.67\%;\,45.29\%]}$	Errado
SPO 1X0 BAH	$46.23\%_{[23.06\%;\,69.98\%]}$	$24.20\%_{[16.99\%;\,31.44\%]}$	$29.56\%_{[11.80\%;\,54.22\%]}$	Certo
VIT 1X1 CRU	$27.80\%_{[11.97\%;\,49.19\%]}$	$27.72\%_{[19.63\%;\,36.35\%]}$	$44.47\%_{[23.51\%;\,66.96\%]}$	Errado
ACG 1X1 CHA	$26.97\%_{[10.18\%;\,48.84\%]}$	$24.65\%_{[16.59\%;\,32.34\%]}$	$48.37\%_{[25.38\%;\,72.49\%]}$	Errado
SAN 1X0 GRE	$31.81\%_{[15.04\%;\,52.88\%]}$	$30.81\%_{[23.10\%;\,39.91\%]}$	$37.36\%_{[19.29\%;\,59.09\%]}$	Errado
CAM 3X0 CFC	$35.35\%_{[16.55\%;\,57.89\%]}$	$26.89\%_{[20.10\%;\,34.56\%]}$	$37.75\%_{[18.37\%;\,60.51\%]}$	Errado
CAP 3X1 VAS	$45.88\%_{[24.72\%;\,68.06\%]}$	$27.44\%_{[19.39\%;\,35.63\%]}$	$26.66\%_{[11.28\%;\,47.52\%]}$	Certo
FLU 2X0 PON	$57.14\%_{[35.07\%;\,78.91\%]}$	$25.69\%_{[15.07\%;\,35.56\%]}$	$17.16\%_{[5.45\%;\,34.96\%]}$	Certo
AVA 2X1 PAL	$16.34\%_{[5.45\%;\,32.47\%]}$	$25.36\%_{[14.86\%;\,35.54\%]}$	$58.28\%_{[37.36\%;\,78.79\%]}$	Errado

5.4 Modelo dinâmico (MD)

A seguir são apresentados os intervalos de 95% de credibilidade das amostras obtidas dos fatores de ataque, defesa e casa da rodada 35:

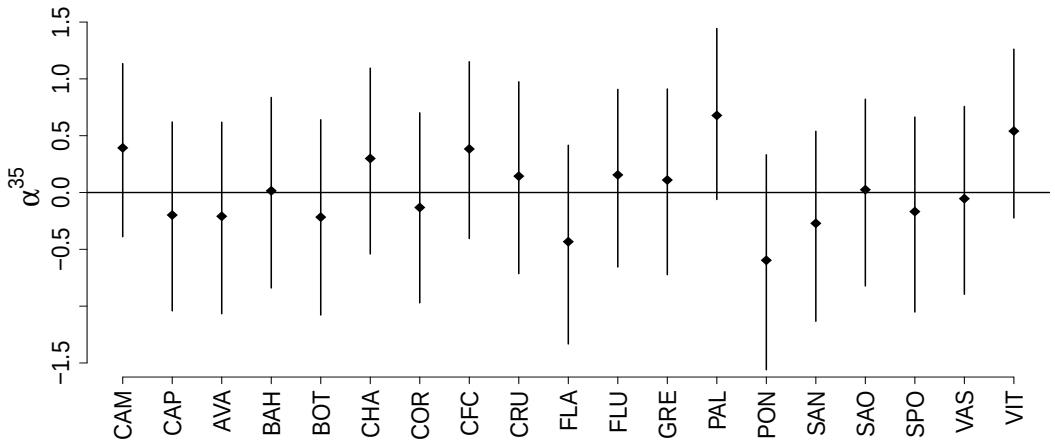


Figura 5.5: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de ataque do modelo MD.

Destaca-se na Figura 5.5 que os times que obtiveram melhores médias referentes aos fatores de ataque foram Palmeiras-SP (0.6784), Vitória-BA (0.5407) e Atlético-MG (0.3929), enquanto Ponte Preta-SP (−0.5965), Flamengo-RJ (−0.4328) e Santos-SP (−0.2714) obtiveram as menores médias das equipes.

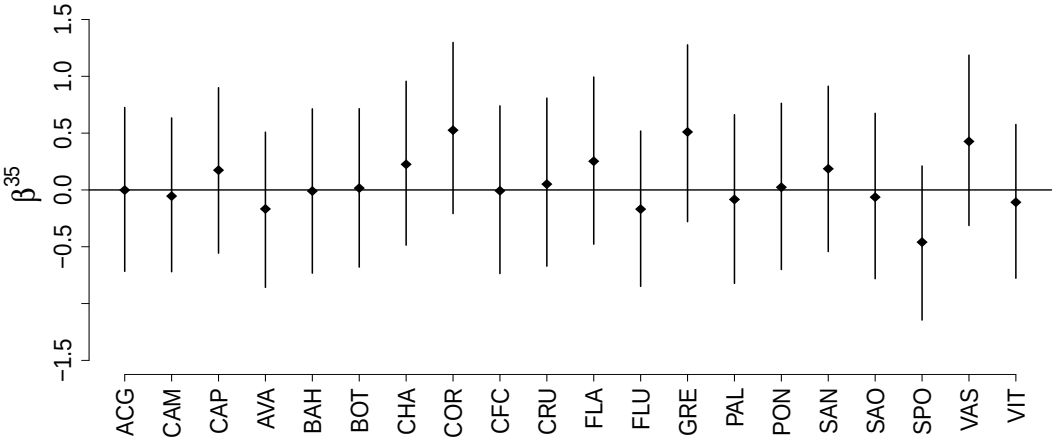


Figura 5.6: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de defesa do modelo MD.

Observando a Figura 5.6 percebe-se que o time que obteve melhor média (0.5259) dos fatores de defesa foi o Corinthians-SP, seguido de Grêmio-RS (0.5102) e Vasco da Gama-RJ (0.4269) e as piores médias foram do Sport-PE (−0.4594), Fluminense-RJ (−0.1689)

e Avaí (−0.1672).

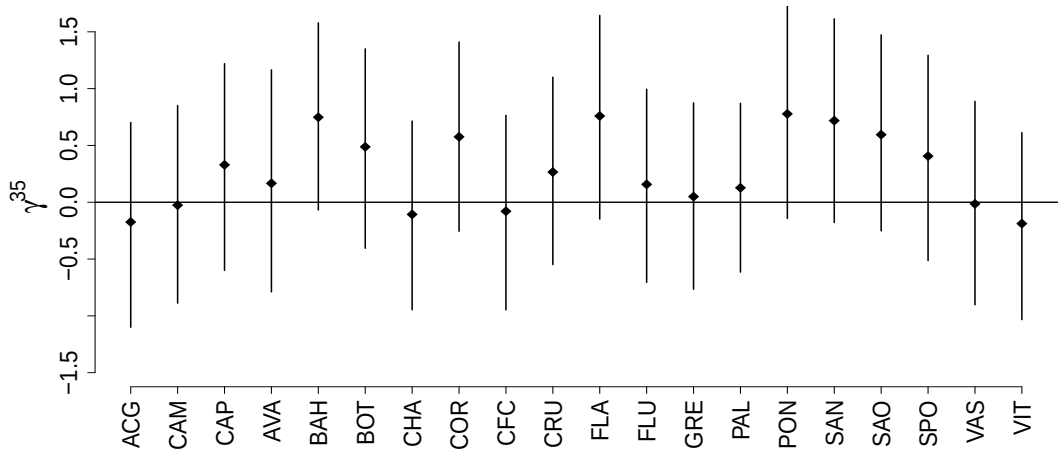


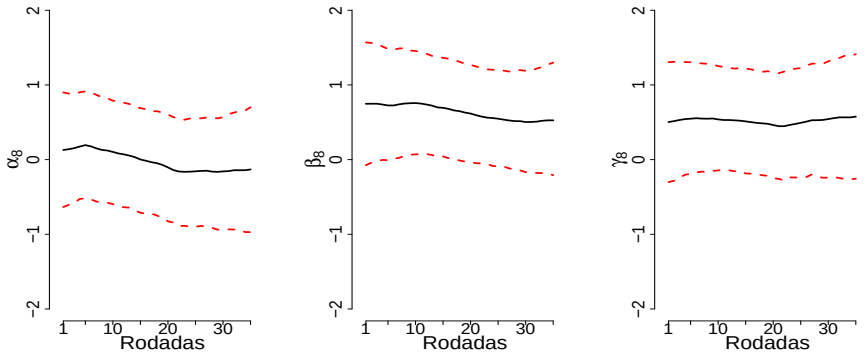
Figura 5.7: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores casa do modelo MD.

Por fim na Figura 5.7 nota-se que as equipes Ponte Preta-SP (0.7776), Flamengo-RJ (0.7591) e Bahia-BA (0.7482) obtiveram melhores médias dos fatores casa e Vitória-BA (−0.1881), Atlético-GO (−0.1741) e Chapecoense-SC (−0.1067) obtiveram as menores.

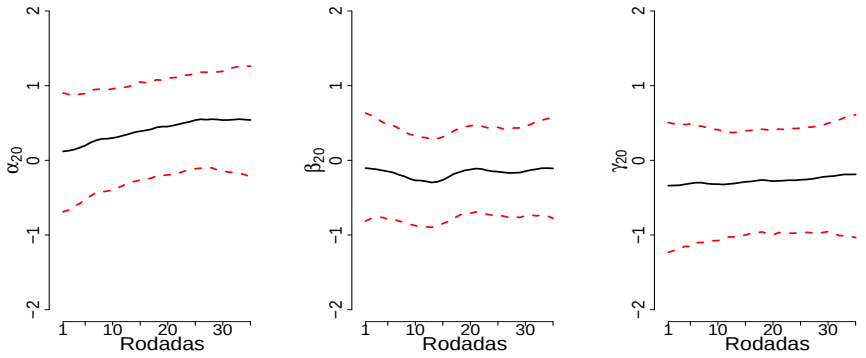
Observe que alguns fatores médios obtidos da rodada 35 do modelo MD diferiram dos fatores estimados do modelo apresentado na Seção 5.3. Destaca-se que o Flamengo-RJ obteve um dos piores fatores de ataque das equipes e o Vasco da Gama-RJ obteve um dos melhores fatores de defesa. Vale ressaltar que o Flamengo-RJ disputou outro torneio ao longo das rodadas finais do campeonato brasileiro, o que provocou uma mudança de foco na equipe carioca. O Vasco da Gama-RJ contratou um novo treinador no decorrer do campeonato. O novo técnico conseguiu montar um sólido sistema defensivo e a equipe começou a sofrer poucos gols. Isso pode explicar as diferentes estimativas encontradas nos dois modelos uma vez que, diferente do modelo ME, o modelo MD pressupõe a evolução dos fatores ao longo das rodadas.

Para avaliar a evolução dos fatores dos times foram selecionados dentre os clubes equipes que apresentaram uma evolução mais significativa em relação aos seus fatores. Vale destacar que alguns fatores de determinadas equipes não apresentaram uma evolução significativa o longo das rodadas. Desse modo, a Figura 5.8 ilustra a evolução dos fatores

de ataque, defesa e casa do Corinthians-SP e Vitória-BA ao longo das rodadas:



(a) Corinthians-SP.

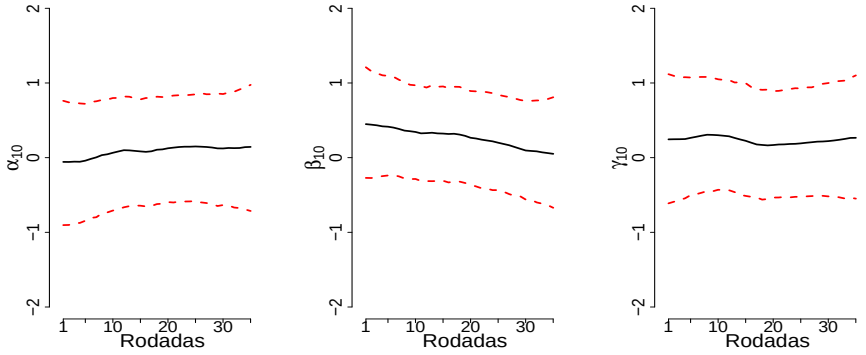


(b) Vitória-BA.

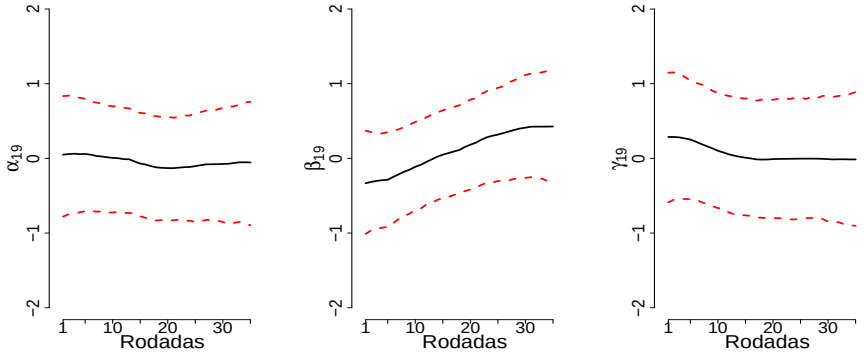
Figura 5.8: Médias *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de ataque (esquerda), defesa (centro) e campo (direita) das equipes do Corinthians-SP (a) e Vitória-BA (b) ao longo das rodadas do modelo MD.

Da Figura 5.8 destaca-se que as estimativas dos fatores de ataque e defesa do Corinthians apresentaram uma queda ao longo das rodadas, enquanto que a estimativa do fator de ataque do Vitória-BA apresentou um crescimento. Além disso nota-se que a estimativa do fator de defesa do time baiano apresentou uma oscilação ao longo das partidas. Com relação aos fatores referentes ao mando de campo de ambas as equipes, nota-se que eles não apresentaram alterações ao longo das rodadas, permanecendo quase estáticos.

A seguir, na Figura 5.9, os fatores de ataque, defesa e casa do Cruzeiro-MG e Vasco da Gama-RJ são apresentados:



(a) Cruzeiro-MG.



(b) Vasco da Gama-RJ.

Figura 5.9: Médias *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de ataque (esquerda), defesa (centro) e campo (direita) das equipes do Cruzeiro-MG (a) e Vasco da Gama-RJ (b) ao longo das rodadas do modelo MD.

Analisando a Figura 5.9 nota-se que a estimativa média do fator de defesa do Cruzeiro-MG teve uma queda ao longo das rodadas, enquanto que a do Vasco da Gama-RJ teve um considerável acréscimo. Além disso, o fator médio referente ao mando de campo da equipe carioca teve uma queda ao longo das partidas, podendo ter sido motivada pela perda de alguns mandos de campo que o time sofreu. Percebe-se que a modelagem dinâmica conseguiu captar a evolução de alguns fatores ao longo das rodadas desses times.

O processo para obtenção das amostras dos números médios de gols das partidas pode ser descrito da seguinte forma: suponha que o time i jogará em casa contra o time j na rodada $t+1$. Obtém-se uma amostra, como descrito anteriormente, das distribuições *a posteriori* dos parâmetros usando os dados até a rodada t, ou seja, obteve-se uma

amostra da distribuição *a posteriori* $\boldsymbol{\theta}^t|D_t$. Para cada elemento obtido amostra-se um elemento da distribuição:

$$\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t \sim NM\left(\boldsymbol{\theta}^t, \mathbf{W}\right) \tag{5.6}$$

Para cada elemento da amostra calcula-se λ_i^{t+1} e λ_j^{t+1}:

$$\lambda_i^{t+1} = exp\left\{\alpha_i^{t+1} - \beta_j^{t+1} + \gamma_i^{t+1}\right\} \tag{5.7}$$

$$\lambda_j^{t+1} = exp\left\{\alpha_j^{t+1} - \beta_i^{t+1}\right\} \tag{5.8}$$

A Tabela 5.4 ilustra as probabilidades médias obtidas da rodada 36 com seus respectivos intervalos de 95% credibilidade:

Tabela 5.4: Médias *a posteriori* e respectivos intervalos de 95% de credibilidade *a posteriori* das probabilidades de vitória, empate e derrota para as partidas da 36ª rodada do modelo MD

Rodada 36	Vitória	Empate	Derrota	Verificação do prognóstico
FLA 3X0 COR	$37.34\%_{[9.12\%;\,76.96\%]}$	$33.92\%_{[14.94\%;\,53.81\%]}$	$28.73\%_{[5.63\%;\,64.68\%]}$	Certo
SAO 0X0 BOT	$58.68\%_{[17.29\%;\,95.94\%]}$	$21.01\%_{[2.92\%;\,39.16\%]}$	$20.30\%_{[0.97\%;\,58.55\%]}$	Errado
SPO 1X0 BAH	$33.24\%_{[4.28\%;\,78.92\%]}$	$21.51\%_{[6.83\%;\,35.99\%]}$	$45.24\%_{[7.95\%;\,86.95\%]}$	Errado
VIT 1X1 CRU	$40.19\%_{[7.21\%;\,86.03\%]}$	$23.16\%_{[7.80\%;\,38.15\%]}$	$36.64\%_{[5.05\%;\,79.34\%]}$	Errado
ACG 1X1 CHA	$20.66\%_{[2.74\%;\,52.97\%]}$	$26.64\%_{[8.31\%;\,44.70\%]}$	$52.69\%_{[19.22\%;\,88.31\%]}$	Errado
SAN 1X0 GRE	$36.04\%_{[7.82\%;\,78.34\%]}$	$29.51\%_{[12.08\%;\,47.86\%]}$	$34.44\%_{[6.80\%;\,72.80\%]}$	Certo
CAM 3X0 CFC	$38.17\%_{[5.67\%;\,86.30\%]}$	$21.15\%_{[6.42\%;\,34.98\%]}$	$40.67\%_{[5.40\%;\,83.56\%]}$	Errado
CAP 3X1 VAS	$32.24\%_{[6.97\%;\,71.92\%]}$	$33.35\%_{[15.27\%;\,52.50\%]}$	$34.39\%_{[7.82\%;\,71.03\%]}$	Errado
FLU 2X0 PON	$52.81\%_{[16.41\%;\,91.73\%]}$	$26.72\%_{[6.09\%;\,47.79\%]}$	$20.46\%_{[1.78\%;\,54.37\%]}$	Certo
AVA 2X1 PAL	$20.07\%_{[0.75\%;\,64.51\%]}$	$17.14\%_{[2.13\%;\,32.00\%]}$	$62.78\%_{[18.11\%;\,97.11\%]}$	Errado

A seguir na Tabela 5.5 e na Figura 5.10 serão apresentadas algumas estatísticas obtidas em relação a variância σ^2 da equação de evolução do modelo e o histograma da amostra obtida:

Tabela 5.5: Resumo do ajuste da variância σ^2 de evolução dos estados do MD

Parâmetro	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
σ^2	0.0099	0.0096	0.0022	0.0064	0.0150

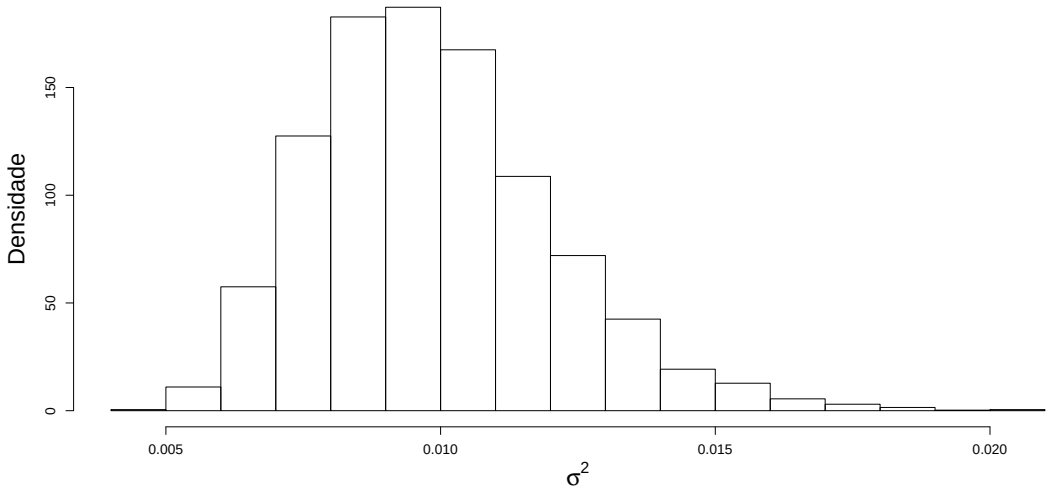


Figura 5.10: Histograma da variância σ^2 de evolução dos estados do MD.

Note que a variância σ^2 de evolução dos estados obteve média 0.0099. Isso indica que há pouca variação na maioria dos fatores dos times ao longo das rodadas.

5.5 Modelo dinâmico com coeficientes auto-regressivos de evolução: MD1 e MD2

A seguir na Tabela 5.6 e Figura 5.11 serão apresentadas algumas estatísticas obtidas em relação aos coeficientes auto-regressivos das equações de evolução e seus respectivos histogramas do modelo MD1:

Tabela 5.6: Resumo do ajuste dos coeficientes auto-regressivos da equação de evolução do MD1

Parâmetro	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
ϕ_α	0.0494	0.0429	0.0355	0.0020	0.1351
ϕ_β	0.0504	0.0439	0.0370	0.0021	0.1365
ϕ_γ	0.0760	0.0700	0.0496	0.0042	0.1904

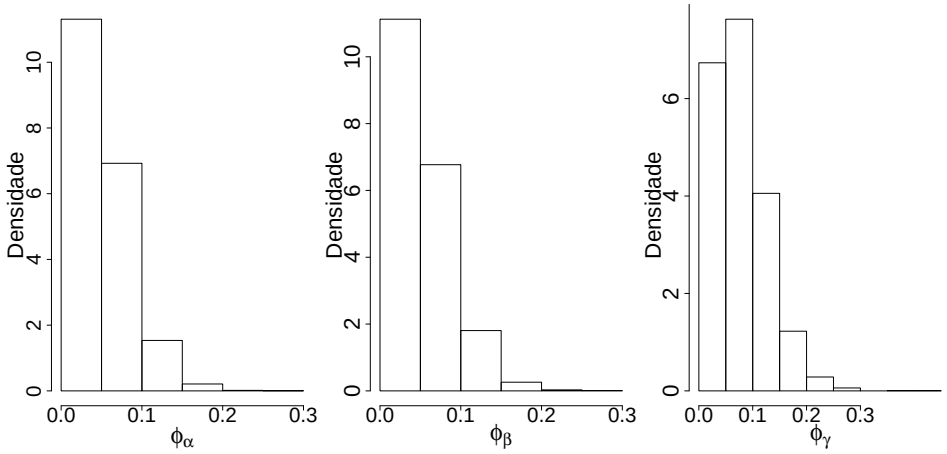


Figura 5.11: Histogramas dos coeficientes auto-regressivos do MD1.

Tanto as médias quanto as medianas dos coeficientes auto-regressivos obtiveram valores muito baixos, indicando que os fatores $\boldsymbol{\theta}^{t-1}$ não influenciam muito os fatores $\boldsymbol{\theta}^t$, ou seja, que não há uma relação muito forte de dependência dos fatores de ataque, defesa e casa de uma rodada para a outra. Como consequência, as estimativas desses fatores ficam todas próximas do zero.

A seguir, na Tabela 5.7, serão apresentadas algumas estatísticas do modelo MD2 referentes aos coeficientes auto-regressivos das equações de evolução dos fatores de ataque ($\phi_{.\alpha}$), defesa ($\phi_{.\beta}$) e casa ($\phi_{.\gamma}$):

Tabela 5.7: Resumo do ajuste dos coeficientes auto-regressivos da equação de evolução do MD2

Parâmetro	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
$\phi_{1\alpha}$	0.0425	0.0345	0.0343	0.0016	0.1264
$\phi_{2\alpha}$	0.0248	0.0204	0.0197	0.0009	0.0728
$\phi_{1\beta}$	0.0501	0.0430	0.0367	0.0026	0.1396
$\phi_{2\beta}$	0.0287	0.0246	0.0212	0.0012	0.0796
$\phi_{1\gamma}$	0.0702	0.0622	0.0490	0.0036	0.1831
$\phi_{2\gamma}$	0.0397	0.0333	0.0309	0.0016	0.1166

As figuras 5.12, 5.13 e 5.14 retratam os histogramas dos coeficientes auto-regressivos do modelo MD2:

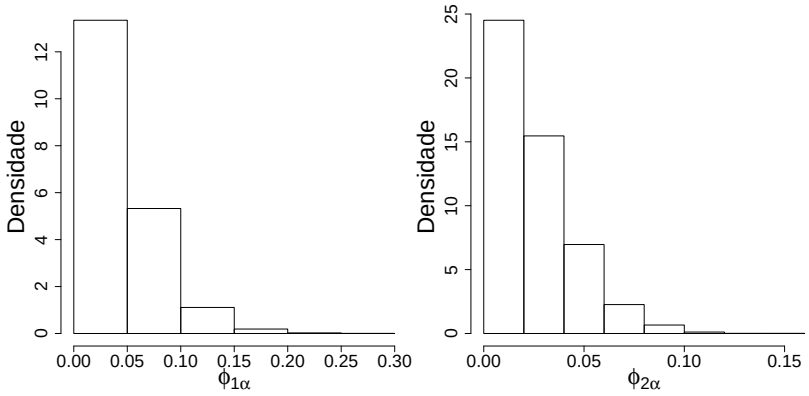


Figura 5.12: Histogramas dos coeficientes auto-regressivos $\phi_{.\alpha}$ do MD2.

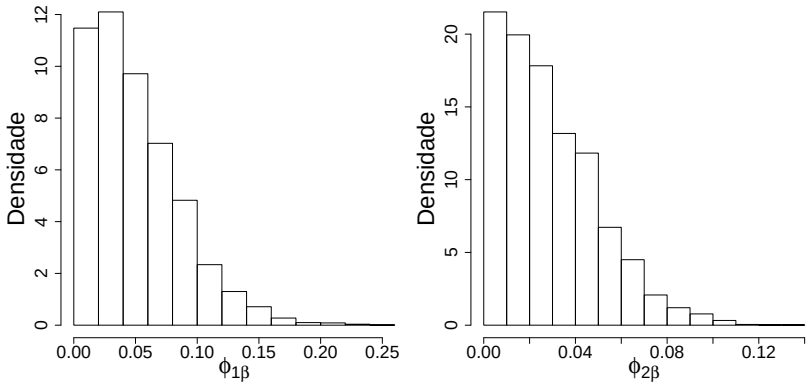


Figura 5.13: Histogramas dos coeficientes auto-regressivos $\phi_{.\beta}$ do MD2.

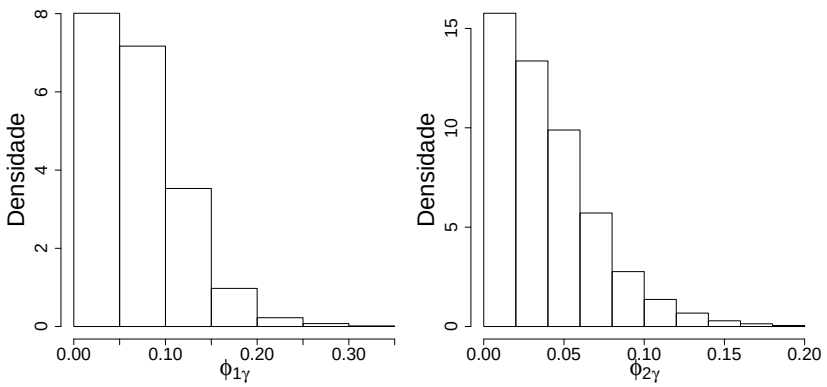


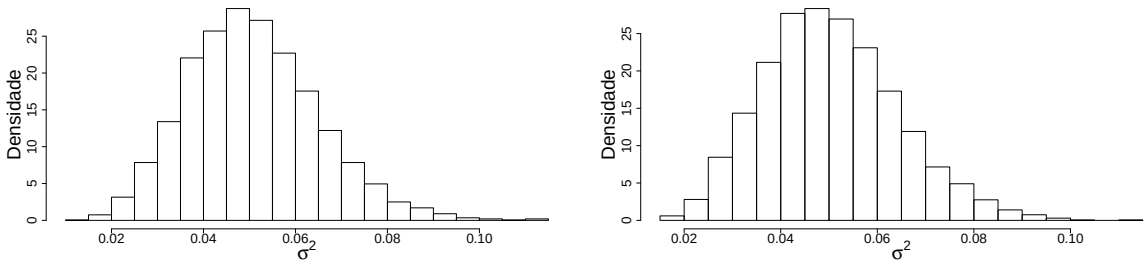
Figura 5.14: Histogramas dos coeficientes auto-regressivos $\phi_{.\gamma}$ do MD2.

Semelhante ao MD1 tanto as médias quanto as medianas dos coeficientes auto-regressivos tiveram valores muito baixos, indicando que os fatores $\boldsymbol{\theta}^{t-1}$ e $\boldsymbol{\theta}^{t-2}$ não influenciam muito os fatores θ^t.

A seguir na Tabela 5.8 e Figura 5.15 serão apresentadas algumas estatísticas obtidas em relação a variância σ^2 da equação de evolução do modelo e o histograma da amostra obtida:

Tabela 5.8: Resumo do ajuste da variância σ^2 de evolução dos estados do MD1 e MD2

σ^2	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
MD1	0.0508	0.0497	0.0142	0.0260	0.0814
MD2	0.0504	0.0495	0.0138	0.0260	0.0805



(a) MD1. (b) MD2.

Figura 5.15: Histograma da variância σ^2 de evolução dos estados do MD1 (a) e MD2(b).

Semelhante ao modelo da Seção 5.4, a variância média obtida foi baixa. Isso indica que

há pouca variação dos fatores ao longo das rodadas. Além disso, em ambos os modelos, tanto as medianas quanto as médias dos fatores ficaram próximas de zero, fazendo com que as previsões de vitória, empate e derrota ficassem aproximadamente iguais.

5.6 Modelo dinâmico com fatores estáticos e com coeficientes auto-regressivos de evolução: MDEST1 e MDEST2

A seguir na Tabela 5.9 e na Figura 5.16 serão apresentadas algumas estatísticas obtidas referentes aos coeficientes auto-regressivos das equações de evolução com seus respectivos histogramas do modelo MDEST1:

Tabela 5.9: Resumo do ajuste dos coeficientes auto-regressivos do MDEST1

Parâmetro	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
ϕ_α	0.0328	0.0272	0.02600	0.0011	0.0949
ϕ_β	0.0338	0.0279	0.0266	0.0013	0.0968
ϕ_γ	0.0379	0.0317	0.0292	0.0013	0.1063

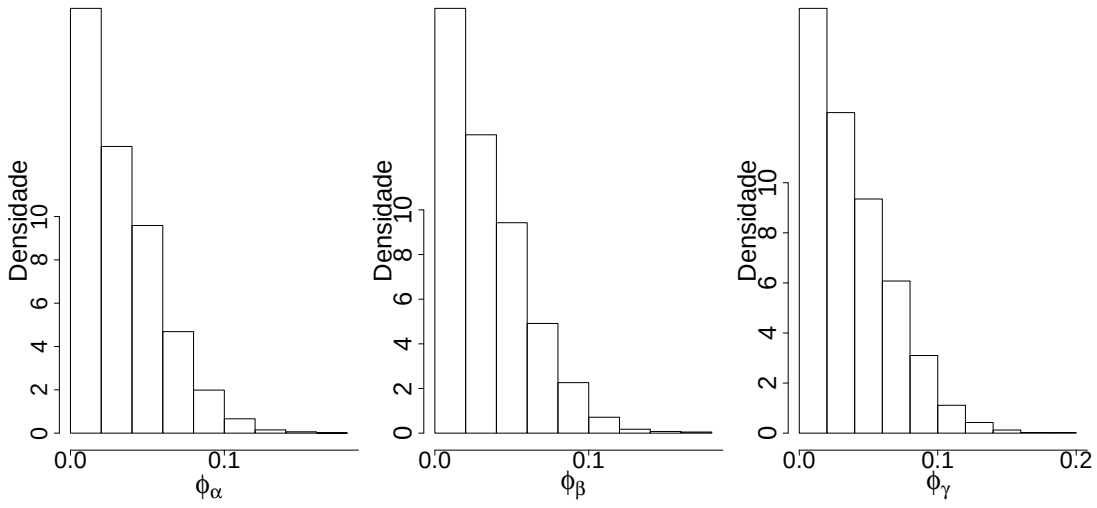


Figura 5.16: Histogramas dos coeficientes auto-regressivos do MDEST1.

Abaixo é apresentado na Tabela 5.10 algumas estatísticas obtidas em relação aos coeficientes auto-regressivos das equações de evolução do modelo MDEST2:

Tabela 5.10: Resumo do ajuste dos coeficientes auto-regressivos da equação de evolução do MDEST2

Parâmetro	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
$\phi_{1\alpha}$	0.0276	0.0227	0.0230	0.0008	0.0864
$\phi_{2\alpha}$	0.0188	0.0144	0.0161	0.0006	0.0600
$\phi_{1\beta}$	0.0332	0.0271	0.0261	0.0014	0.0964
$\phi_{2\beta}$	0.0219	0.0184	0.0172	0.0008	0.0641
$\phi_{1\gamma}$	0.0331	0.0262	0.0275	0.0010	0.1026
$\phi_{2\gamma}$	0.0232	0.0183	0.0197	0.0007	0.0720

A Figura 5.17 corresponde aos histogramas dos coeficientes auto-regressivos $\phi_{.\alpha}$ do modelo MD2:

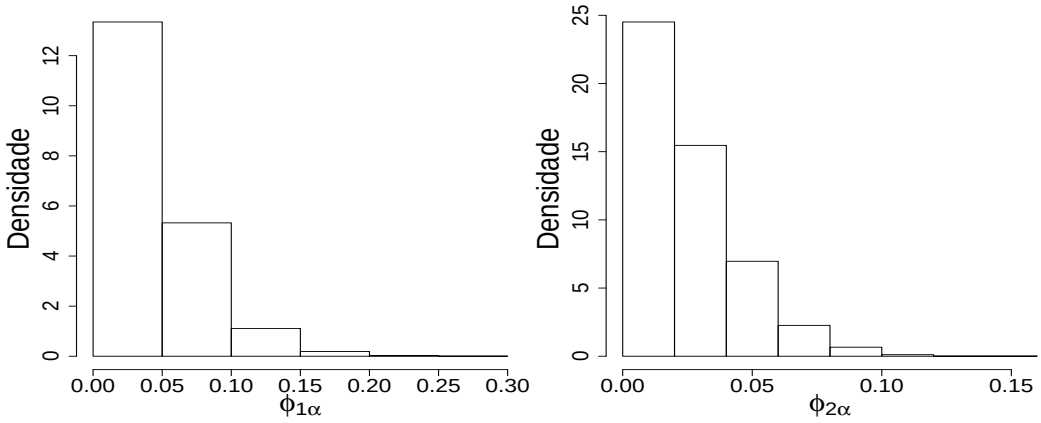


Figura 5.17: Histogramas dos coeficientes auto-regressivos $\phi_{.\alpha}$ do MDEST2.

A seguir a Figura 5.18 corresponde aos histogramas dos coeficientes auto-regressivos $\phi_{.\beta}$ do modelo MD2:

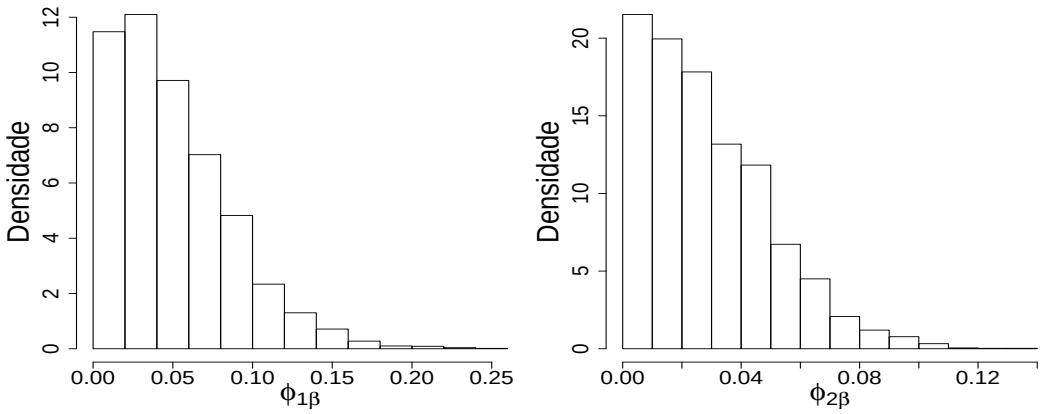


Figura 5.18: Histogramas dos coeficientes auto-regressivos $\phi_{.\beta}$ do MDEST2.

A Figura 5.19, que está sendo apresentada abaixo, corresponde aos histogramas dos coeficientes auto-regressivos $\phi_{.\gamma}$ do modelo MD2:

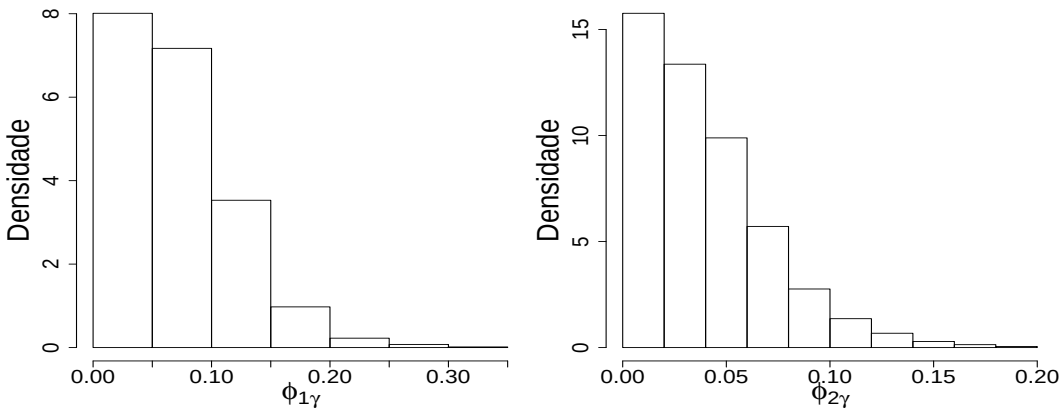


Figura 5.19: Histogramas dos coeficientes auto-regressivos $\phi_{.\gamma}$ do MDEST2.

Note que em ambos os modelos os coeficientes auto-regressivos obtiveram médias muito baixas. Isso significa que na prática os modelos estão quase se aproximando ao modelo estático, uma vez que a parte correspondente à evolução dos fatores foi estimada quase nula.

São apresentados na Figura 5.20 os intervalos de 95% de credibilidade das amostras obtidas dos fatores de ataque dos modelos ME, MDEST1 e MDEST2:

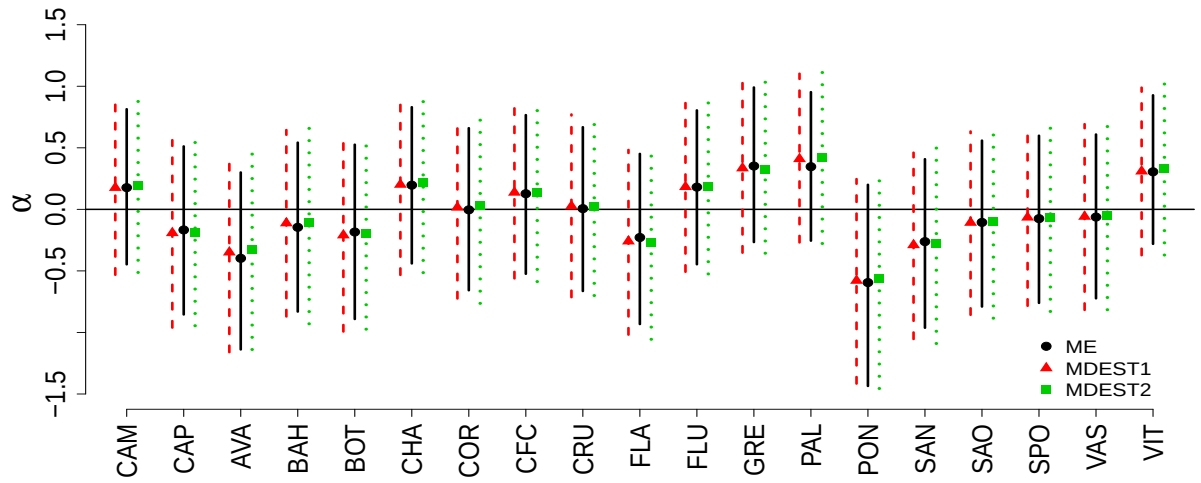


Figura 5.20: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de ataque do modelo MD, MDEST1 e MDEST2.

A seguir na Figura 5.21 são apresentados os intervalos de 95% de credibilidade das amostras obtidas dos fatores de defesa da rodada 35 dos modelos ME, MDEST1 e MDEST2:

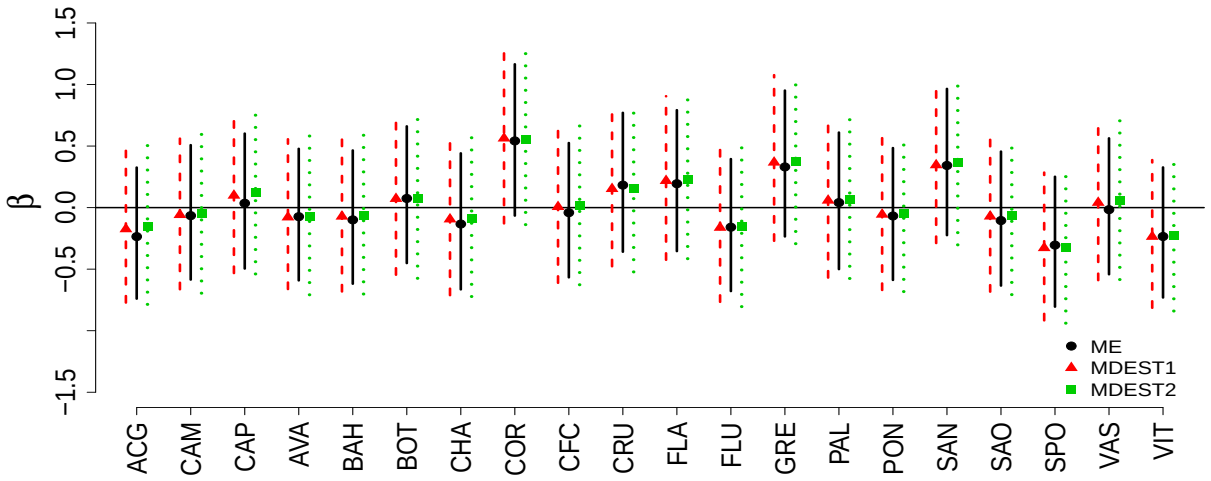


Figura 5.21: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores de defesa dos modelos MD, MDEST1 e MDEST2.

Na Figura 5.22 são apresentados os intervalos de 95% de credibilidade das amostras obtidas dos fatores casa da rodada 35 dos modelos ME, MDEST1 e MDEST2:

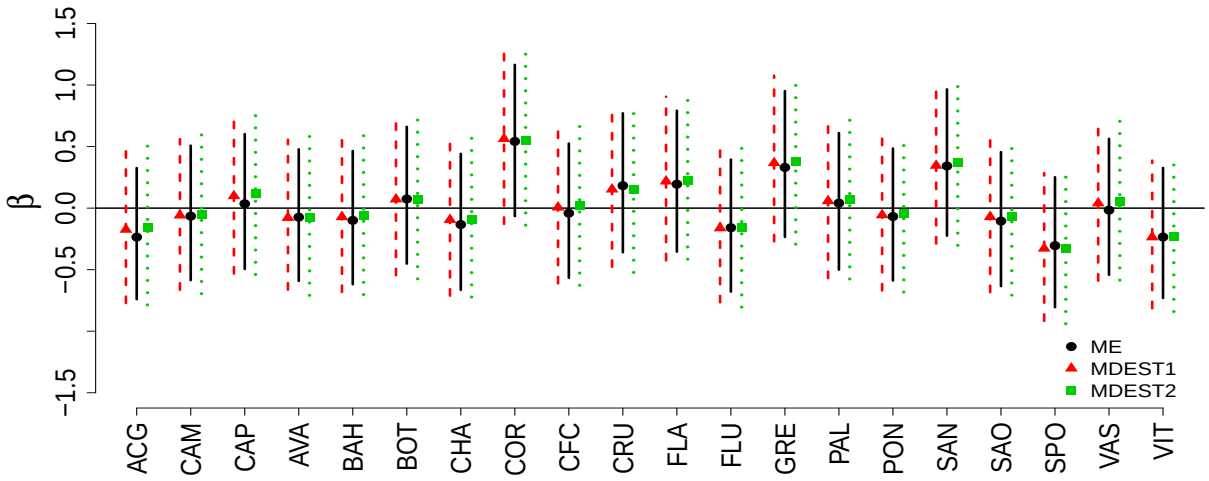
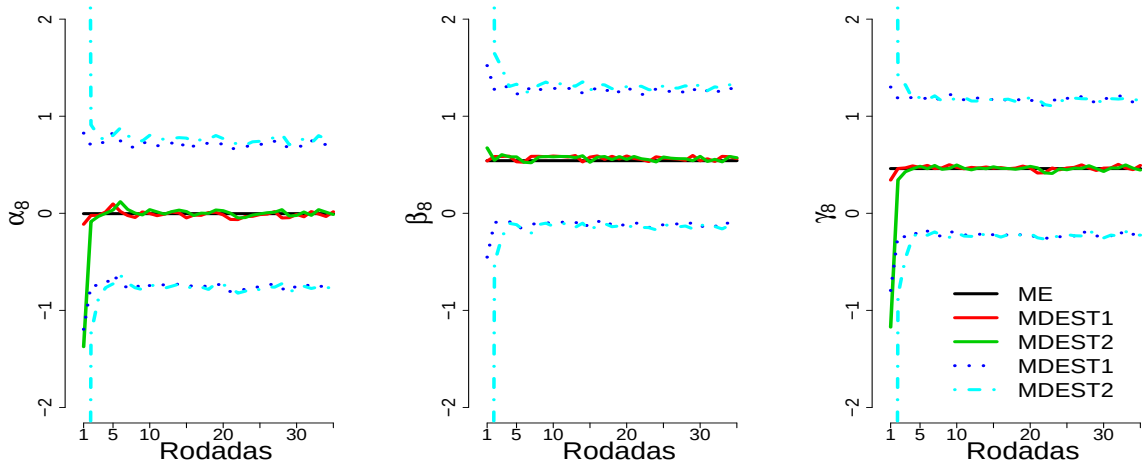


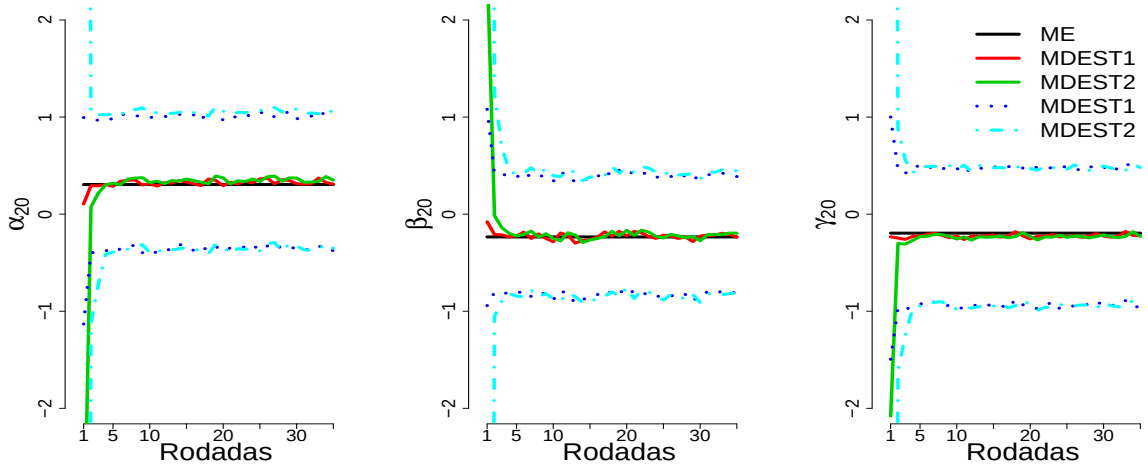
Figura 5.22: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores casa do modelo MD, MDEST1 e MDEST2..

Note que, diferente dos modelos MD1 e MD2, agora as médias dos fatores não estão próximas de zero e estão muito próximas das médias obtidas no modelo ME, indicando que os fatores não evoluem no tempo dinamicamente, ou seja, que os fatores estão estáticos ao longo das rodadas.

Para avaliar a evolução dos fatores dos times foram selecionados os mesmos clubes do MD a fim de comparar se os times apresentaram evolução ao longo das rodadas. Desse modo, na Figura 5.23 será ilustrada a evolução dos fatores de ataque, defesa e casa do Corinthians-SP e Vitória-BA; na Figura 5.24 os fatores do Cruzeiro-MG e Vasco da Gama-RJ ao longo das rodadas:

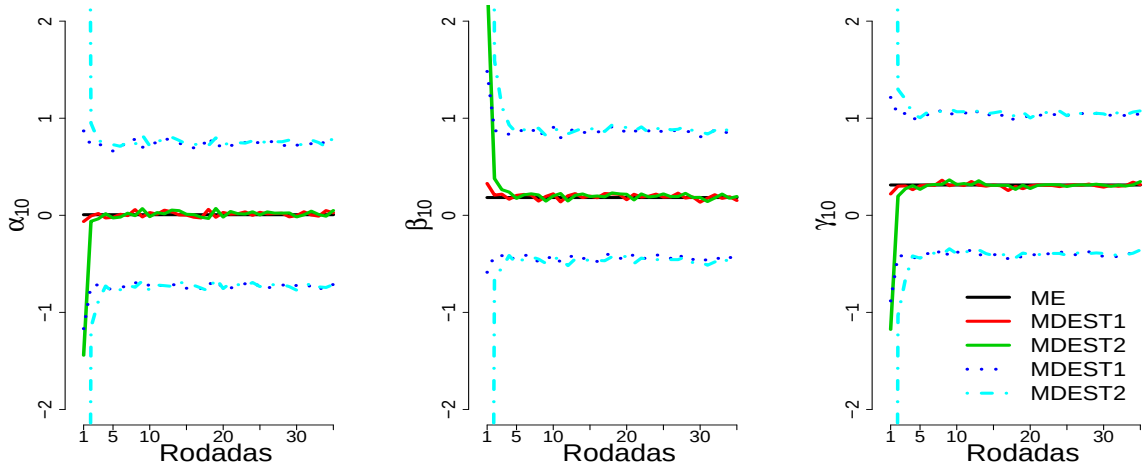


(a) Fatores do Corinthians-SP.

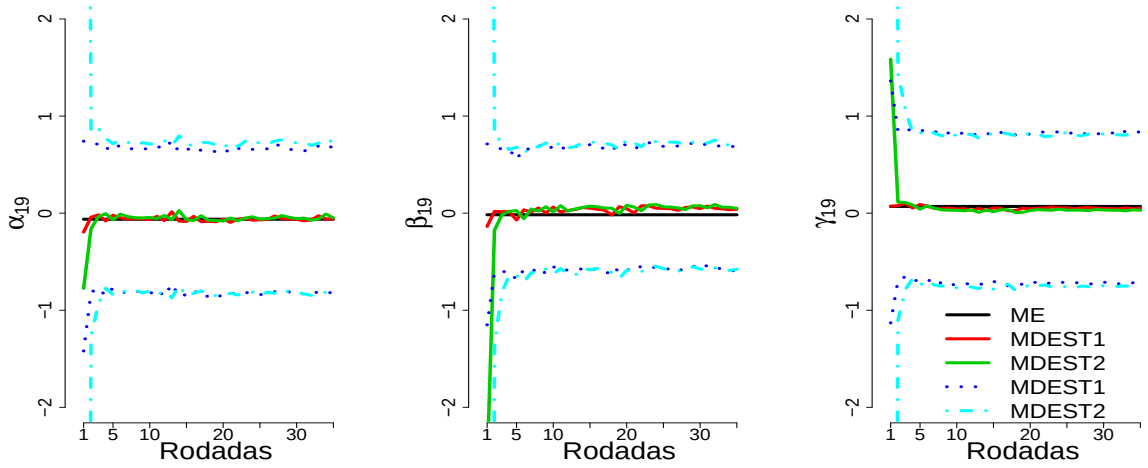


(b) Fatores do Vitória-BA.

Figura 5.23: Médias *a posteriori* (linhas cheias) e intervalos de 95% de credibilidade *a posteriori* (linhas tracejadas) dos fatores de ataque (esquerda), defesa (centro) e campo (direita) das equipes do Corinthians-SP (a) e Vitória-BA (b) ao longo das rodadas dos modelos MD, MDEST1 e MDEST2.



(a) Fatores do Cruzeiro-MG.



(b) Fatores do Vasco da Gama-RJ.

Figura 5.24: Médias *a posteriori* (linhas cheias) e intervalos de 95% de credibilidade *a posteriori* (linhas tracejadas) dos fatores de ataque (esquerda), defesa (centro) e campo (direita) das equipes do Cruzeiro-MG (a) e Vasco da Gama-RJ (b) ao longo das rodadas dos modelos MD, MDEST1 e MDEST2.

Analisando as figuras 5.23 e 5.24, nota-se que a evolução dos fatores médios obtidos nos modelos MDEST1 e MDEST2 ficaram em torno das médias do modelo ME, reforçando o indício que a maioria dos fatores das equipes participantes não evoluem dinamicamente ao longo das rodadas.

O processo para obtenção das amostras dos números médios de gols das partidas dos modelos MDEST1 e MDEST2 podem ser descritos da seguinte forma: suponha que o time i jogará em casa contra o time j na rodada $t+1$. Obtém-se uma amostra das distribuições *a posteriori* dos parâmetros usando os dados até a rodada t, ou seja, obteve-se uma amostra da distribuição *a posteriori* $\boldsymbol{\theta}^t|D_t$. Para cada elemento obtido amostra-se um elemento da distribuição:

$$\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t \sim NM\left(\boldsymbol{\kappa} + \boldsymbol{\phi}\boldsymbol{\theta}^t, \mathbf{W}\right), \tag{5.9}$$

referente ao modelo MDEST1;

$$\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t \sim NM\left(\boldsymbol{\kappa} + \boldsymbol{\phi}_1\boldsymbol{\theta}^t + \boldsymbol{\phi}_2\boldsymbol{\theta}^{t-1}, \mathbf{W}\right), \tag{5.10}$$

referente ao modelo MDEST2. O restante do procedimento é similar ao do modelo apresentado na Seção 5.4.

A Tabela 5.11 a seguir ilustra as probabilidades médias obtidas da rodada 36 com seus respectivos intervalos de 95% de credibilidade do modelo MDEST1:

Tabela 5.11: Médias *a posteriori* e respectivos intervalos de 95% de credibilidade *a posteriori* das probabilidades de vitória, empate e derrota para as partidas da 36ª rodada do modelo MDEST1

Rodada 36	Vitória	Empate	Derrota	Verificação do prognóstico
FLA 3X0 COR	$35.00\%_{[12.74\%;\,66.52\%]}$	$33.18\%_{[19.65\%;\,47.29\%]}$	$31.81\%_{[10.28\%;\,60.49\%]}$	Certo
SAO 0X0 BOT	$49.50\%_{[19.59\%;\,83.72\%]}$	$25.37\%_{[10.80\%;\,38.13\%]}$	$25.11\%_{[5.48\%;\,55.22\%]}$	Errado
SPO 1X0 BAH	$46.44\%_{[16.38\%;\,80.58\%]}$	$23.94\%_{[11.83\%;\,35.35\%]}$	$29.61\%_{[6.78\%;\,62.50\%]}$	Certo
VIT 1X1 CRU	$28.26\%_{[7.08\%;\,59.14\%]}$	$27.18\%_{[14.68\%;\,39.71\%]}$	$44.55\%_{[17.04\%;\,76.55\%]}$	Errado
ACG 1X1 CHA	$27.84\%_{[6.91\%;\,58.17\%]}$	$24.72\%_{[13.04\%;\,35.41\%]}$	$47.43\%_{[18.76\%;\,79.35\%]}$	Errado
SAN 1X0 GRE	$32.92\%_{[11.40\%;\,63.98\%]}$	$30.86\%_{[19.27\%;\,44.14\%]}$	$36.20\%_{[12.67\%;\,64.83\%]}$	Errado
CAM 3X0 CFC	$34.62\%_{[10.36\%;\,68.27\%]}$	$26.40\%_{[14.38\%;\,38.32\%]}$	$38.96\%_{[12.25\%;\,71.40\%]}$	Errado
CAP 3X1 VAS	$45.04\%_{[17.89\%;\,77.64\%]}$	$28.16\%_{[14.85\%;\,41.53\%]}$	$26.79\%_{[7.16\%;\,54.99\%]}$	Certo
FLU 2X0 PON	$54.95\%_{[24.53\%;\,86.27\%]}$	$26.69\%_{[9.88\%;\,41.88\%]}$	$18.35\%_{[3.33\%;\,43.35\%]}$	Certo
AVA 2X1 PAL	$17.03\%_{[3.30\%;\,43.06\%]}$	$24.94\%_{[10.45\%;\,39.40\%]}$	$58.01\%_{[27.72\%;\,85.49\%]}$	Errado

Observe que as probabilidades médias obtidas foram muito parecidas com as do ME, entretanto os comprimentos dos intervalos de credibilidade foram maiores.

A Tabela 5.12 retrata as probabilidades médias obtidas da rodada 36 com seus respectivos intervalos de 95% de credibilidade do modelo MDEST2:

Tabela 5.12: Médias *a posteriori* e respectivos intervalos de 95% de credibilidade *a posteriori* das probabilidades de vitória, empate e derrota para as partidas da 36ª rodada do modelo MDEST2

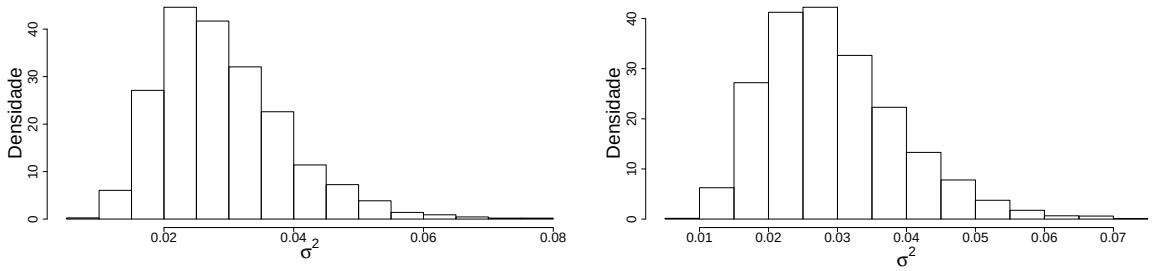
Rodada 36	Vitória	Empate	Derrota	Verificação do prognóstico
FLA 3X0 COR	$35.50\%_{[12.63\%;\ 66.46\%]}$	$32.76\%_{[19.50\%;\ 47.48\%]}$	$31.72\%_{[10.30\%;\ 60.38\%]}$	Certo
SAO 0X0 BOT	$49.65\%_{[19.93\%;\ 83.69\%]}$	$25.25\%_{[10.65\%;\ 38.00\%]}$	$25.08\%_{[5.15\%;\ 55.57\%]}$	Errado
SPO 1X0 BAH	$46.27\%_{[15.51\%;\ 80.40\%]}$	$24.02\%_{[11.71\%;\ 35.68\%]}$	$29.70\%_{[6.78\%;\ 62.30\%]}$	Certo
VIT 1X1 CRU	$28.62\%_{[7.45\%;\ 59.36\%]}$	$27.12\%_{[14.24\%;\ 39.90\%]}$	$44.24\%_{[16.44\%;\ 76.01\%]}$	Errado
ACG 1X1 CHA	$28.14\%_{[6.91\%;\ 58.91\%]}$	$24.74\%_{[12.40\%;\ 36.13\%]}$	$47.10\%_{[17.96\%;\ 80.24\%]}$	Errado
SAN 1X0 GRE	$33.76\%_{[11.63\%;\ 65.07\%]}$	$31.03\%_{[19.16\%;\ 44.58\%]}$	$35.19\%_{[12.38\%;\ 63.55\%]}$	Errado
CAM 3X0 CFC	$34.92\%_{[10.46\%;\ 68.84\%]}$	$26.42\%_{[15.13\%;\ 38.16\%]}$	$38.65\%_{[11.91\%;\ 71.17\%]}$	Errado
CAP 3X1 VAS	$45.28\%_{[17.31\%;\ 79.19\%]}$	$28.17\%_{[13.80\%;\ 41.58\%]}$	$26.54\%_{[6.45\%;\ 55.68\%]}$	Certo
FLU 2X0 PON	$54.77\%_{[24.00\%;\ 87.22\%]}$	$26.65\%_{[9.27\%;\ 42.07\%]}$	$18.57\%_{[3.10\%;\ 44.45\%]}$	Certo
AVA 2X1 PAL	$16.96\%_{[3.15\%;\ 41.88\%]}$	$24.85\%_{[9.65\%;\ 39.31\%]}$	$58.18\%_{[28.64\%;\ 86.79\%]}$	Errado

Observe que as probabilidades médias *a posteriori* e os intervalos de 95% de credibilidade obtidos da rodada 36 de ambos os modelos ficaram muito próximas.

Vale destacar que semelhante aos modelo das seções 5.4 e 5.5, a variância média obtida dos modelos MDEST1 e MDEST2 foram baixas. A título de ilustração a seguir na Tabela 5.13 e Figura 5.25 serão apresentadas algumas estatísticas obtidas em relação a variância σ^2 da equação de evolução dos modelos MDEST1 e MDEST2 e os seus respectivos histogramas das amostras obtidas:

Tabela 5.13: Resumo do ajuste da variância σ^2 de evolução dos estados dos modelos MDEST1 e MDEST2

σ^2	Média	Mediana	Erro-padrão	$Q_{2,5\%}$	$Q_{97,5\%}$
MDEST1	0.0291	0.0276	0.0099	0.0145	0.0519
MDEST2	0.0293	0.0278	0.0098	0.0145	0.0519



(a) MDEST1. (b) MDEST2.

Figura 5.25: Histograma da variância σ^2 de evolução dos estados do MDEST1 (a) e MDEST2(b).

5.7 Modelo hierárquico estático (MHE)

As figuras a seguir são os intervalos de 95% de credibilidade das amostras obtidas dos coeficientes referentes a número de finalizações, escanteios, faltas e cartões:

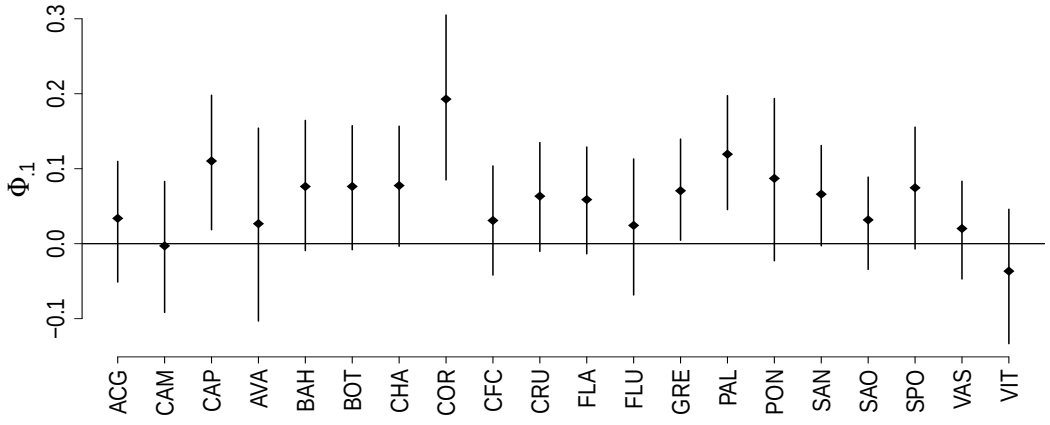


Figura 5.26: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* referentes ao número de finalizações do MHE.

Analisando a Figura 5.26 destaca-se que as maiores médias dos coeficientes obtidos referentes a finalização foram Corinthians-SP (0.1928), Palmeiras-SP (0.1194) e Atlético-PR (0.1103) e as menores foram Vitória-BA (−0.0366), Atlético-MG (−0.0029) e Vasco da Gama (0.0202).

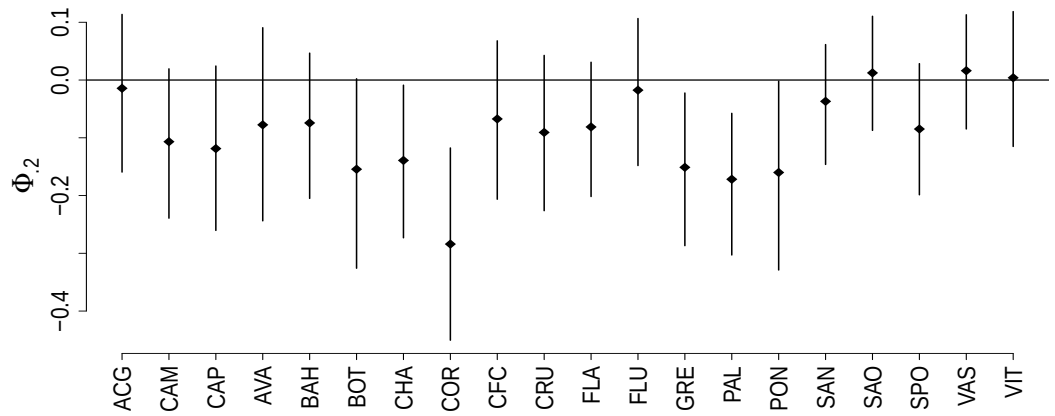


Figura 5.27: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* referentes ao número de escanteios do MHE.

Analisando a Figura 5.27 referentes aos coeficientes dos escanteios as maiores médias foram do Vasco da Gama-RJ (0.0161), São-Paulo-SP (0.0123) e Vitória-BA (0.0039) e as menores foram Corinthians-SP (−0.2840), Palmeiras-SP (−0.1718) e Ponte-Preta-SP (−0.1601).

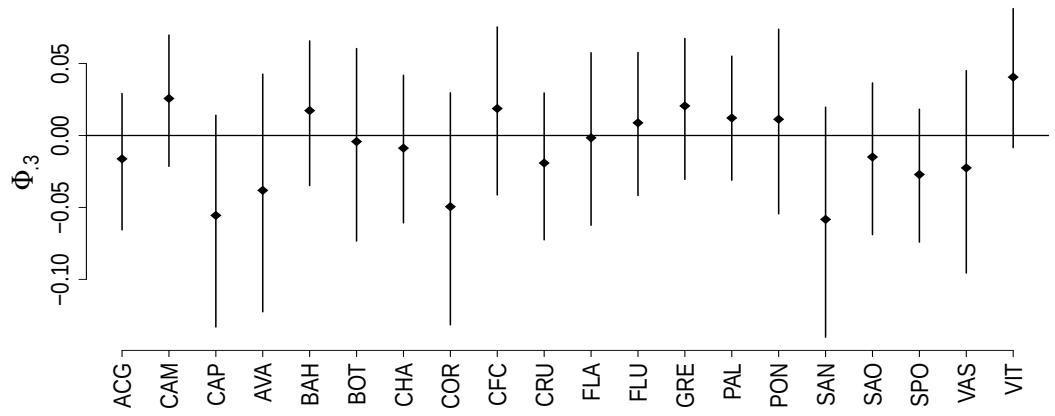


Figura 5.28: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* referentes ao número de faltas do MHE.

Nota-se na Figura 5.28 que as maiores médias referentes aos coeficientes das faltas as foram Vitória-BA (0.0406), Atlético-MG (0.0257) e Grêmio-RS (0.02054) e as menores foram Santos-SP (−0.0583), Atlético-PR (−0.0555) e Corinthians-SP (−0.0495).

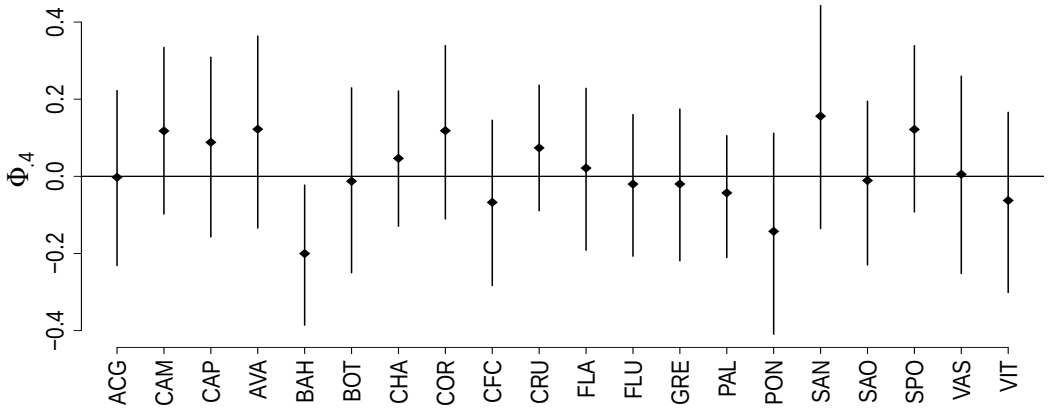


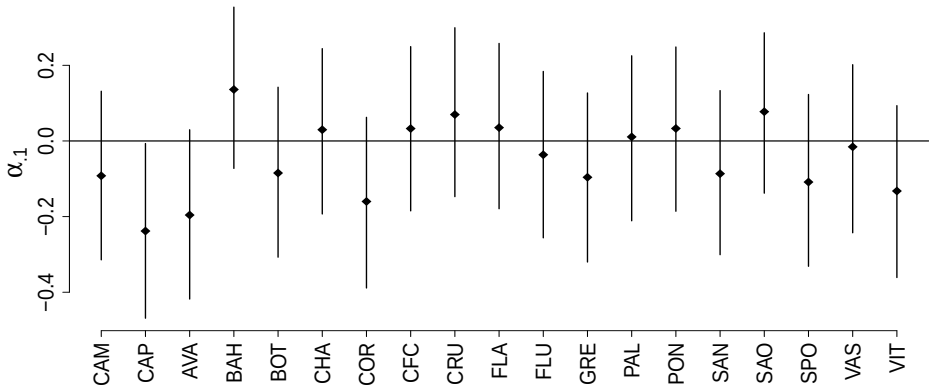
Figura 5.29: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* referentes ao número de cartões do MHE.

Destaca-se da Figura 5.29 que as maiores médias dos coeficientes obtidos referentes aos cartões as maiores médias foram Santos-SP (0.1560), Avaí-SC (0.1222) e Sport-PE (0.1216) e as menores foram Bahia-BA (−0.2003), Ponte-Preta-SP (−0.1427) e Coritiba-PR (−0.0675).

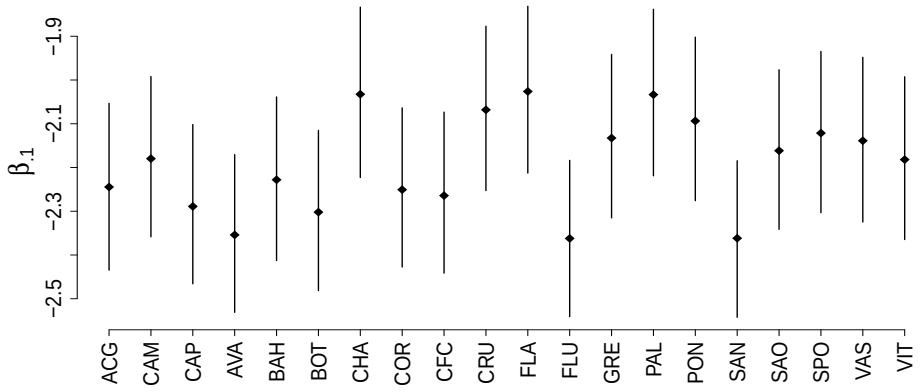
Analisando as figuras anteriores pode-se conhecer determinadas características das equipes participantes do campeonato. Destaca-se que o Corinthians-SP, campeão do torneio, obteve o melhor coeficiente referentes ao número de finalizações. Em contrapartida a equipe obteve o menor coeficiente médio referente aos escanteios. Conclui-se que o número de finalizações da equipe paulista contribui em maior peso e que o número de escanteios contribuem em menor peso na sua média de gols.

O Vitória-BA foi uma das equipes do campeonato que obteve bons resultados fora de casa e sua principal característica era o jogo em contra-ataque, tendo poucas finalizações durante os jogos. Em contrapartida a equipe não conseguia obter bons resultados dentro de casa, uma vez que os clubes mandantes tendem a ter um número maior de finalizações, característica oposta ao seu estilo de jogo. O coeficiente médio obtido foi o menor de todos (−0.0366), ou seja, o coeficiente conseguiu captar essa característica da equipe.

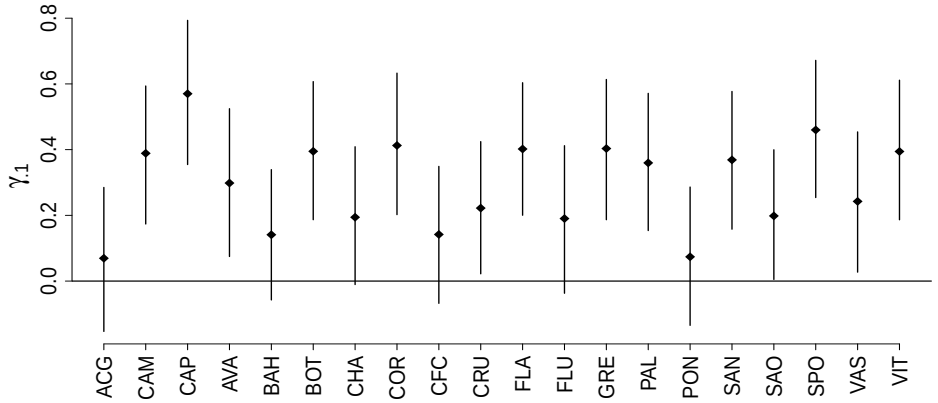
A seguir na Figura 5.30 serão apresentados os intervalos de 95% de credibilidade dos fatores $\alpha_{.1}$, $\beta_{.1}$ e $\gamma_{.1}$ referentes à variável número de finalizações:



(a) Fatores $\alpha_{.1}$.



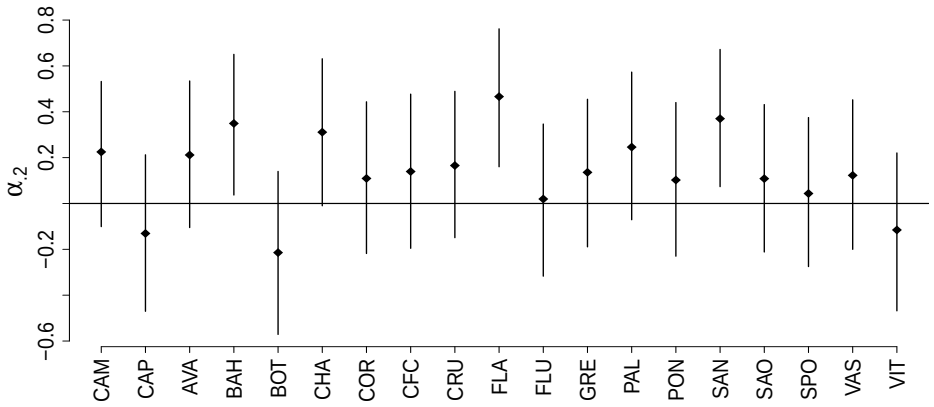
(b) Fatores $\beta_{.1}$.



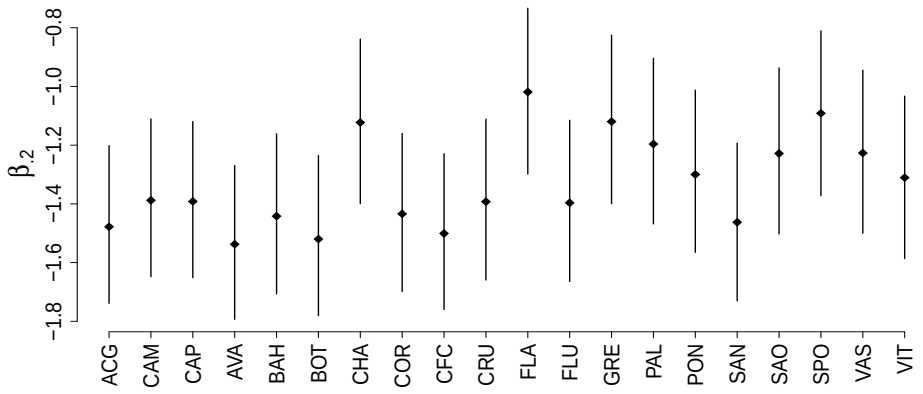
(c) Fatores $\gamma_{.1}$.

Figura 5.30: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores $\alpha_{.1}$, $\beta_{.1}$ e $\gamma_{.1}$ referentes ao número de finalizações do MHE.

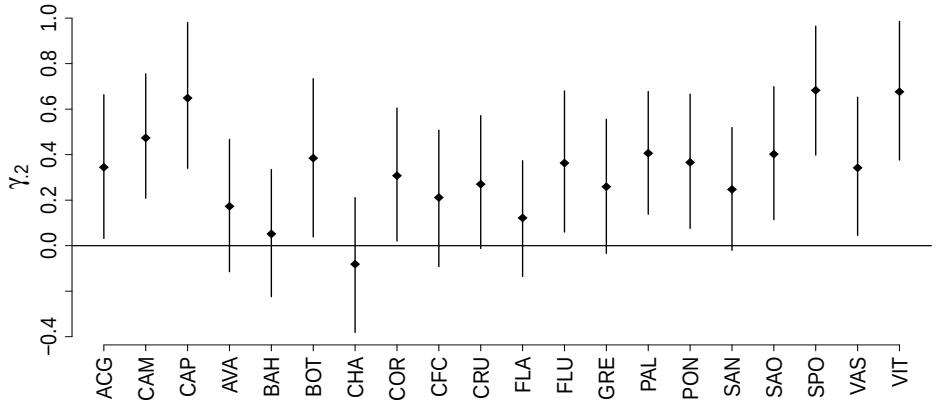
A seguir na Figura 5.31 serão apresentados os intervalos de 95% de credibilidade dos fatores $\alpha_{.2}$, $\beta_{.2}$ e $\gamma_{.2}$ referentes ao número de escanteios:



(a) Fatores $\alpha_{.2}$.



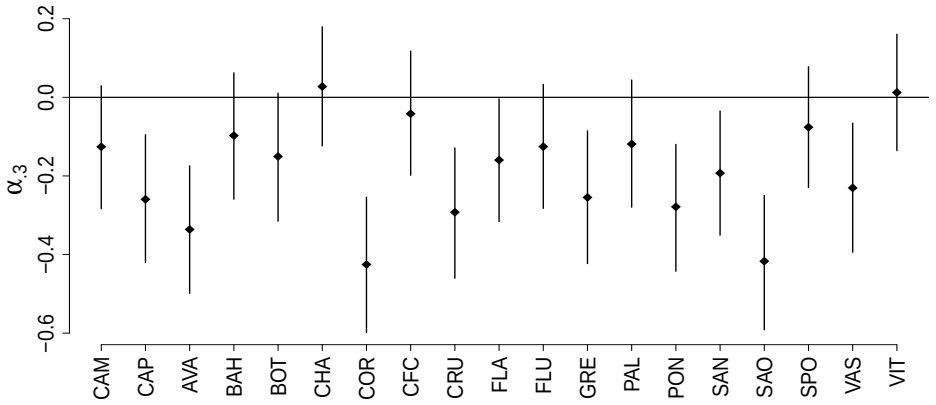
(b) Fatores $\beta_{.2}$.



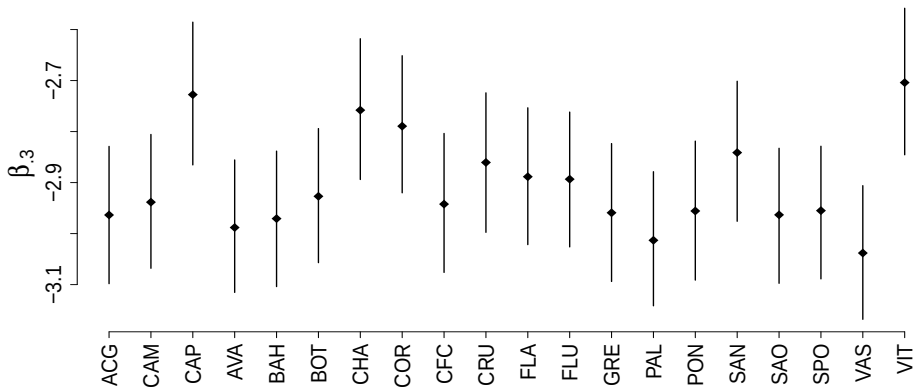
(c) Fatores $\gamma_{.2}$.

Figura 5.31: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores $\alpha_{.2}$, $\beta_{.2}$ e $\gamma_{.2}$ referentes ao número de escanteios do MHE.

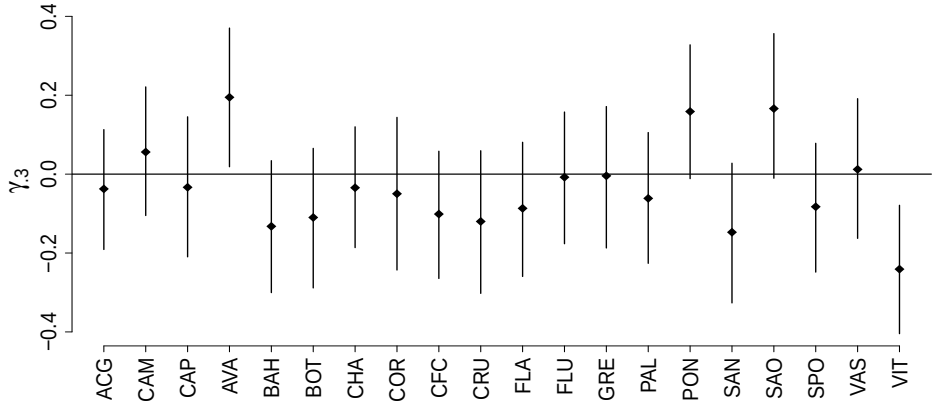
Abaixo na Figura 5.32 serão apresentados os intervalos de 95% de credibilidade dos fatores $\alpha_{.3}$, $\beta_{.3}$ e $\gamma_{.3}$ referentes ao número de faltas:



(a) Fatores $\alpha_{.3}$.



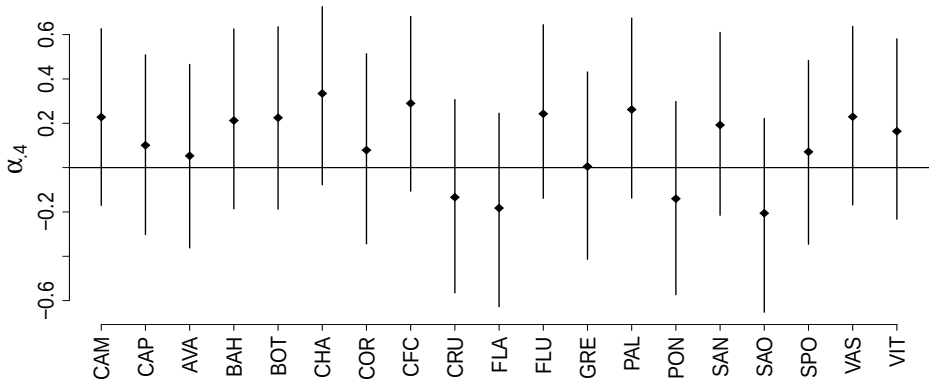
(b) Fatores $\beta_{.3}$.



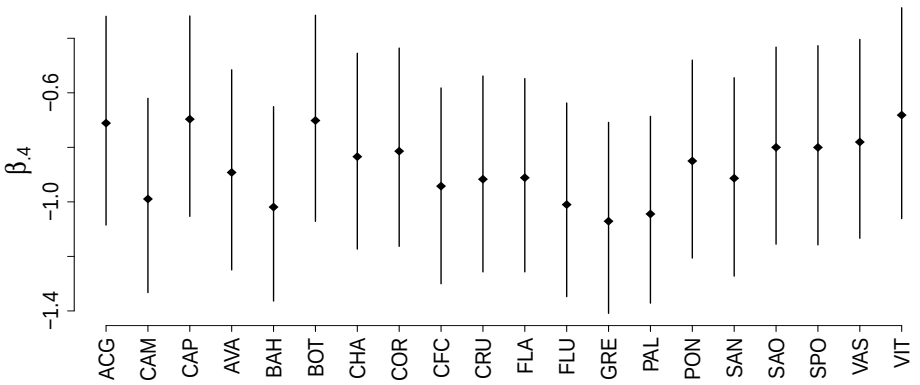
(c) Fatores $\gamma_{.3}$.

Figura 5.32: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores $\alpha_{.3}$, $\beta_{.3}$ e $\gamma_{.3}$ referentes ao número de faltas do MHE.

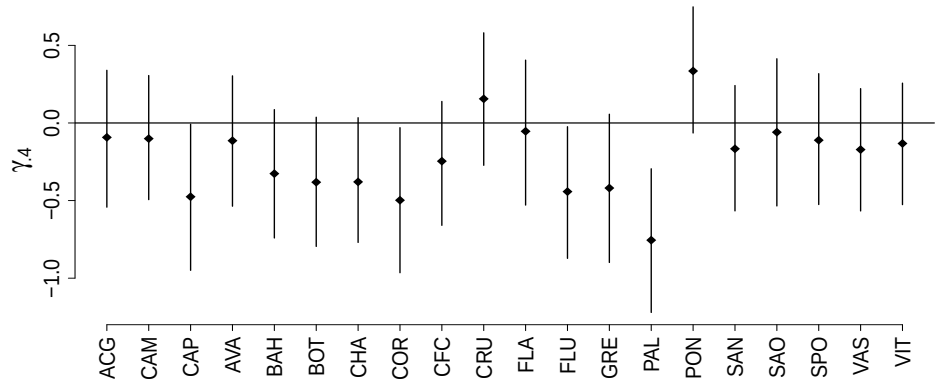
A seguir na Figura 5.33 serão apresentados os intervalos de 95% de credibilidade dos fatores da variável número de cartões:



(a) Fatores $\alpha_{.4}$.



(b) Fatores $\beta_{.4}$.



(c) Fatores $\gamma_{.4}$.

Figura 5.33: Média *a posteriori* e intervalos de 95% de credibilidade *a posteriori* dos fatores $\alpha_{.4}$, $\beta_{.4}$ e $\gamma_{.4}$ referentes ao número de cartões do MHE.

Analisando a Figura 5.30 percebe-se que os times que obtiveram melhores médias dos fatores de $\alpha_{.1}$ foram Bahia-BA (0.1360), São-Paulo-SP (0.0774) e Cruzeiro-MG (0.0698), enquanto Atlético-PR (-0.2381), Avaí-SC (-0.1959) e Corinthians-SP (-0.1600) obtiveram as menores médias das equipes. Com relação aos fatores $\beta_{.1}$ nota-se que o Flamengo-RJ obteve melhor média (-2.0262) em relação as demais equipes seguido de Chapecoense-SC (-2.0325) e Palmeiras-SP (-2.0334) e Fluminense-RJ (-2.3622), Santos-SP (-2.3617) e Avaí-SC (-2.3541) obtiveram as menores médias. Por fim com relação aos fatores $\gamma_{.1}$ nota-se que equipes como Atlético-PR (0.5701), Sport-PE (0.4599) e Corinthians-SP (0.4126) obtiveram maiores médias e Atlético-GO (0.0694), Ponte-Preta-SP (0.0738) e Bahia-BA (0.1410) obtiveram as menores médias. Destaca-se que o Corinthians-SP (equipe que obteve o melhor coeficiente da regressão referente à finalização) obteve o pior fator médio $\alpha_{.1}$. Entretanto, em relação aos fatores $\gamma_{.1}$ obteve um dos melhores fatores médios. Conclui-se que a equipe paulista obtém um número de finalizações mais expressivas jogando como mandante e por consequência aumenta o seu número médio de gols, crescendo as chances de fazer gols e vencer as partidas disputadas em casa.

Com relação aos fatores referentes ao número de escanteios (Figura 5.31) nota-se que os times que obtiveram melhores médias dos fatores de $\alpha_{.2}$ foram Flamengo-RJ (0.4661), Santos-SP (0.3699) e Bahia-BA (0.3490) enquanto Botafogo-RJ (-0.2143), Atlético-PR (-0.1308) e Vitória-BA (-0.1155) obtiveram as menores médias das equipes. Com relação aos fatores $\beta_{.2}$ nota-se que o Flamengo-RJ obteve melhor média (-1.019) em relação as demais equipes seguido de Sport-PE (-1.0913) e Grêmio (-1.1197) e Avaí-SC (-1.5376), Botafogo-RJ (-1.5199) e Coritiba-PR (-1.5008) obtiveram as menores médias. Finalmente com relação aos fatores $\gamma_{.2}$ nota-se que equipes como Sport-PE (0.6827), Vitória-BA (0.6762) e Atlético-PR (0.6487) obtiveram maiores médias e Chapecoense-SC (-0.0813), Bahia-BA (0.0517) e Flamengo-RJ (0.1220) obtiveram as menores médias. Destaca-se que o Flamengo-RJ obteve o melhor fator médio $\alpha_{.2}$ e $\beta_{.2}$, e um dos piores fatores médios $\gamma_{.2}$, ou seja, a equipe tende a ter um número de escanteios grandes, jogando como mandante ou visitante e contribui em menor escala para o número de escanteios das equipes adversárias. Mas note que o time carioca não teve um

dos melhores coeficientes médios da regressão referentes aos escanteios. Conclui-se que um dos modos da equipe aumentar o seu número médio de gols seria ter um número grande de escanteios.

Destaca-se na Figura 5.32 que os times que obtiveram melhores médias dos fatores de $\alpha_{.3}$ foram Chapecoense-SC (0.0275), Vitória-BA (0.0122) e Coritiba-PR (-0.0417) enquanto Corinthians-SP (-0.4254), São-Paulo-SP (-0.4170) e Avaí-SC (-0.3361) obtiveram as menores médias das equipes. Com relação aos fatores $\beta_{.3}$ nota-se que o Vitória-BA obteve melhor média (-2.7040) em relação as demais equipes seguido de Atlético-PR (-2.7275) e Chapecoense-SC (-2.7580) e Vasco da Gama-RJ (-3.0381), Palmeiras-SP (-3.0131) e Avaí-SC (-2.9881) obtiveram as menores médias. Por fim com relação aos fatores $\gamma_{.3}$ nota-se que equipes como Avaí-SC (0.1948), São-Paulo-SP (0.1661) e Ponte-Preta-SP (0.1587) obtiveram maiores médias e Vitória-BA (-0.2409), Santos-SP (-0.1473) e Bahia-BA (-0.1324) obtiveram as menores médias. É interessante destacar que os fatores que mais contribuem para aumentar o número médio de faltas das equipes são $\beta_{.3}$, referentes aos adversários que elas enfrentam nas partidas, ou seja, conforme o adversário enfrentado a equipe tende a cometar mais faltas ou não.

Por fim analisando a Figura 5.33 nota-se que os times que obtiveram melhores médias dos fatores de $\alpha_{.4}$ foram Chapecoense-SC (0.3342), Coritiba-PR (0.2898) e Palmeiras-SP (0.2619) enquanto São-Paulo-SP (-0.2057), Flamengo-RJ (-0.1825) e Palmeiras-SP (-0.1403) obtiveram as menores médias das equipes. Note que Com relação aos fatores $\beta_{.4}$ nota-se que o Vitória-BA obteve melhor média (-0.6818) em relação as demais equipes seguido de Atlético-PR (-0.6967) e Botafogo-RJ (-0.7016) e Grêmio-RS (-1.0709), Palmeiras-SP (-1.0444) e Bahia-BA (-1.0190) obtiveram as menores médias. Por último em relação aos fatores $\gamma_{.4}$ nota-se que equipes como Ponte-Preta-SP (0.3348), Cruzeiro-MG (0.1555) e Flamengo-RJ (-0.05370) obtiveram maiores médias e Palmeiras-SP (-0.7552), Corinthians-SP (-0.4979) e Atlético-PR (-0.4754) obtiveram as menores médias.

Uma vez que modelou-se também as variáveis conjuntamente, os passos para prever o número de gols do time mandante e visitante na rodada t (por consequência, o resultado) passam primeiro pela previsão das variáveis nesta mesma rodada. Sendo assim, o processo

para obtenção das amostras dos números médios de gols das partidas pode ser descrito da seguinte forma: suponha que o time i jogará em casa contra o time j na rodada $t+1$. Obtém-se uma amostra, como descrito anteriormente, das distribuições *a posteriori*, tanto dos coeficientes das variáveis quanto dos fatores das variáveis, usando os dados até a rodada t. Para cada elemento amostral de cada variável amostra-se um elemento das distribuições:

$$X_{ik}^{t+1}|\eta_{ik}^{t+1} \sim Poisson(\eta_{ik}^{t+1}) \tag{5.11}$$

$$X_{jk}^{t+1}|\eta_{jk}^{t+1} \sim Poisson(\eta_{jk}^{t+1}) \tag{5.12}$$

onde:

$$\eta_{ik}^{t+1} = exp\{\alpha_{ik} - \beta_{jk} + \gamma_{ik}\} \tag{5.13}$$

$$\eta_{jk}^{t+1} = exp\{\alpha_{jk} - \beta_{ik}\} \tag{5.14}$$

Uma vez obtido as previsões das variáveis número de finalizações, escanteios, faltas e cartões, calula-se o número médio de gols do time i e j:

$$\lambda_{i}^{t+1} = exp\left(\Phi_{i1}x_{i1}^{t+1} + \Phi_{i2}x_{i2}^{t+1} + \Phi_{i3}x_{i3}^{t+1} + \Phi_{i4}x_{i4}^{t+1}\right), \tag{5.15}$$

$$\lambda_{j}^{t+1} = exp\left(\Phi_{j1}x_{j1}^{t+1} + \Phi_{j2}x_{j2}^{t+1} + \Phi_{j3}x_{j3}^{t+1} + \Phi_{j4}x_{j4}^{t+1}\right). \tag{5.16}$$

Uma vez realizado esse processo pode-se fazer previsões para os placares da rodada 36. Note que, diferente dos demais modelos apresentados, nesse é necessário prever as variáveis consideradas. A Tabela 5.14 a seguir ilustra as probabilidades médias obtidas da rodada 36 com seus respectivos intervalos de 95% de credibilidade *a posteriori*:

Tabela 5.14: Médias *a posteriori* e respectivos intervalos de 95% de credibilidade *a posteriori* das probabilidades de vitória, empate e derrota para as partidas da 36ª rodada do modelo MHE

Rodada 36	Vitória	Empate	Derrota	Verificação do prognóstico
FLA 3X0 COR	$39.89\%_{[2.02\%;\ 86.34\%]}$	$23.22\%_{[2.87\%;\ 42.93\%]}$	$36.88\%_{[2.96\%;\ 93.91\%]}$	Certo
SAO 0X0 BOT	$41.80\%_{[11.07\%;\ 75.96\%]}$	$25.28\%_{[12.78\%;\ 37.56\%]}$	$32.91\%_{[6.68\%;\ 73.30\%]}$	Errado
SPO 1X0 BAH	$36.66\%_{[5.02\%;\ 84.37\%]}$	$25.29\%_{[6.51\%;\ 43.37\%]}$	$38.04\%_{[4.68\%;\ 82.59\%]}$	Errado
VIT 1X1 CRU	$34.67\%_{[8.85\%;\ 69.90\%]}$	$27.25\%_{[14.02\%;\ 39.60\%]}$	$38.07\%_{[10.89\%;\ 72.94\%]}$	Errado
ACG 1X1 CHA	$35.77\%_{[10.35\%;\ 65.55\%]}$	$30.17\%_{[15.96\%;\ 44.49\%]}$	$34.05\%_{[7.26\%;\ 71.36\%]}$	Errado
SAN 1X0 GRE	$38.81\%_{[6.62\%;\ 87.98\%]}$	$24.57\%_{[5.93\%;\ 40.09\%]}$	$36.60\%_{[3.62\%;\ 79.55\%]}$	Certo
CAM 3X0 CFC	$30.14\%_{[4.34\%;\ 70.16\%]}$	$28.98\%_{[12.11\%;\ 46.13\%]}$	$40.86\%_{[9.81\%;\ 78.29\%]}$	Errado
CAP 3X1 VAS	$34.60\%_{[6.47\%;\ 80.97\%]}$	$29.04\%_{[10.34\%;\ 44.54\%]}$	$36.34\%_{[6.84\%;\ 68.02\%]}$	Errado
FLU 2X0 PON	$41.02\%_{[6.40\%;\ 76.21\%]}$	$25.18\%_{[8.34\%;\ 38.34\%]}$	$33.79\%_{[5.95\%;\ 84.07\%]}$	Certo
AVA 2X1 PAL	$19.09\%_{[0.62\%;\ 55.30\%]}$	$26.06\%_{[2.06\%;\ 50.09\%]}$	$54.83\%_{[12.34\%;\ 97.09\%]}$	Errado

5.8 Critérios de comparação dos modelos

Será apresentado na Tabela 5.15 as taxas de acertos obtidas dos modelos ME, MD, MDEST1, MDEST2 e MHE:

Tabela 5.15: Comparação dos modelos

Modelos	Taxas de acerto
ME	40.00%
MD	30.00%
MDEST1	40.00%
MDEST2	40.00%
MHE	30.00%

Com a relação a capacidade preditiva dos modelos as taxas de acerto ficaram entre

30% e 40%. Embora pareça baixa, para uma das últimas rodadas do campeonato onde a situação de várias equipes já estavam definidas pode-se considerar uma boa taxa de previsão. Por exemplo, nas partidas Avaí-SC vs Palmeiras-SP e Vitória-BA vs Cruzeiro-MG destaca-se que os clubes mandantes estavam lutando contra o rebaixamento e os visitantes já estavam classificados para Libertadores e não disputavam mais o título, sendo que o Cruzeiro-MG tinha sido campeão da Copa do Brasil, outra competição de destaque no Brasil. Desse modo é natural esperar que o Avaí-SC e o Vitória-BA tivessem um melhor rendimento em comparação aos clubes visitantes, como foi constatado na vitória do Avaí-SC e no empate do Vitória-BA. Além disso em várias ocasiões as probabilidades de vitória e derrota dos clubes mandantes ficaram muito próximas e os intervalos de credibilidades ficaram com comprimentos altos, reforçando o fato que o Campeonato Brasileiro é um torneio muito equilibrado e de difícil previsão.

Para comparar os modelos ajustados anteriormente serão aplicados dois critérios de comparação: o DIC (*Deviance Information Criterion*), proposto por Spiegelhalter et al. (2002) e o RPS (*Ranked Probability Scores*), proposto por Gneiting et al. (2007).

O DIC é pode ser descrito como:

$$DIC = \bar{D} + p_D, \tag{5.17}$$

onde $D(\boldsymbol{\theta}) = -2logL(\mathbf{y}|\boldsymbol{\theta})$ é a distribuição *a posteriori* da *deviance*. Observe que o DIC é composto pela soma de dois termos: o primeiro $\bar{D}$ é a média *a posteriori* da *deviance*, uma medida de adequação do modelo e o $p_D = \bar{D} - D(\bar{\theta})$, onde $\bar{\theta}$ são o conjuntos das médias *a posteriori* dos elementos do vetor paramêtrico $\boldsymbol{\theta}$, uma medida de penalidade do número de parâmetros do modelo. O menor valor DIC obtido dentre os modelos apontará qual foi o melhor modelo ajustado aos dados.

O RPS é especificado por:

$$RPS = E|y_{rep} - y| - \frac{1}{2}E|y_{rep} - \tilde{y}_{rep}|, \tag{5.18}$$

$E|y_{rep} - y| = \frac{1}{n}\sum_{i=1}^{n}|y_{rep}(s_i) - y(s_i)|$, $E|y_{rep} - \bar{y}_{rep}| = \frac{1}{n}\sum_{i=1}^{n}|y_{rep}(s_i) - \tilde{y}_{rep}(s_i)|$, onde $y(s_i)$ são as observações, $y_{rep}(s_i)$ e $\tilde{y}_{rep}(s_i)$ são os valores replicados independentemente da

distribuição preditiva para cada uma das localizações amostradas. O menor valor RPS apontará qual foi o melhor modelo ajustado.

Observe que o primeiro critério apresentado utilizou para o seu cálculo a função de verossimilhança, tanto a verossimilhança média quanto a verossimilhança calculada a partir dos valores dos parâmetros médios e o segundo utilizou a previsão da variável. Nesse sentido pode-se classificar o DIC como um critério de informação e o RPS um critério preditivo.

Os valores obtidos dos modelos são apresentados na Tabela 5.16 abaixo:

Tabela 5.16: Comparação dos modelos

Modelos	$\bar{D}$	p_D	DIC	RPS
ME	1951.27	59.49	2010.76	0.5434
MD	1906.80	120.87	2027.67	0.4916
MD1	1896.28	100.63	1996.91	0.5031
MD2	1895.17	103.65	1998.82	0.4987
MDEST1	1901.40	114.69	2016.09	0.4912
MDEST2	1900.24	117.74	2017.98	0.4903
MHE	1947.79	79.05	2026.84	0.5316

Segundo o critério DIC o modelo que se ajustou melhor aos dados foi o MD1, seguido do MD2. Observe que foram justamente os modelos que apontaram que os clubes estão muito nivelados e que não apresentam grandes diferenças nos seus fatores, em desconcordância aos outros modelos implementados e os dados do próprio campeonato. Isso pode indicar que a maioria dos clubes do campeonato estão no mesmo patamar e que para efeitos de estimação não seja vantajoso destacar as melhores equipes. Além disso, o pior ajuste foi do MD, justamente o modelo que permite a evolução dos fatores ao longo mas sem os coeficientes auto-regressivos. Com relação ao RPS o modelo que se ajustou melhor foi o MDEST2, embora os valores obtidos tenham ficado próximos.

Analisando as taxas de acertos obtidas e o critério de comparação RPS, o modelo que melhor se ajustou aos dados do campeonato foi o MDETS2, modelo proposto na presente dissertação.

Capítulo 6

Conclusões

Neste trabalho foram ajustados modelos para placares de partidas de futebol e para previsão. Dentre os modelos ajustados, alguns foram propostos nesta dissertação: os modelos MD2, MDEST1, MDEST2 e MHE. Todos os modelos propostos podem ser aplicados em outras edições do campeonato ou ainda em outros torneios de pontos corridos. Com relação aos modelos que consideram somente coeficientes auto-regressivos nas equações de evolução (MD1 e MD2), destaca-se que, segundo seus pressupostos, os fatores das equipes não diferiram muito e ficaram próximos de zero. Isso foi provocado pelas estimativas dos coeficientes auto-regressivos terem ficadas próximas de zero, fazendo com que as cadeias obtidas ficassem em torno desse valor. Por consequência as probabilidades de vitória, empate e derrota de todas as partidas da rodada 36 ficaram quase iguais. Esses modelos apontaram que a maioria dos fatores não apresentam uma evolução ao longo do tempo, o que foi confirmado ao ajustar-se os modelos com fatores estáticos e coeficientes auto-regressivos (MDEST1 e MDEST2). Neles as estimativas dos fatores estáticas conseguiram captar as diferenças de forças entre os clubes, diferente dos fatores auto-regressivos, correspondentes a parte dinâmica.

Analisando a tabela final do campeonato, percebe-se que os fatores estimados de alguns modelos se aproximaram com os dados dos clubes. Destaca-se que os modelos estático (ME), dinâmico (MD) e os com fatores estáticos e coeficientes auto-regressivos (MDEST1 e MDEST2) apontaram os fatores de ataque do Palmeiras-SP, Grêmio-RS, Atlético-MG e Vitória-BA como os melhores dentre as equipes. Os times em questão

tiveram as maiores quantidades de gols marcados (61, 55, 52 e 50, nessa ordem). Clubes como Avaí-SC, Ponte-Preta-SP e Atlético-GO obtiveram os piores fatores de ataque e de fato foram as equipes que menos marcaram gols (29, 37 e 38, nessa ordem). O Corinthians-SP obteve o melhor fator de defesa. A equipe paulista sofreu o menor número de gols do campeonato (31). O pior fator defensivo foi do Sport-PE, o clube que sofreu o maior número de gols do campeonato (58). Para finalizar o melhor fator de campo apontado por todos os modelos citados foi o correspondente a Ponte-Preta-SP e o pior foi do Vitória-BA. Vale ressaltar que dos 39 pontos obtidos pela Ponte-Preta-SP, 30 (76.92%) foram jogando como mandante e que dos 43 pontos obtidos pelo Vitória-BA, 14 (32.55%) foram jogando em casa, o pior mandante do campeonato. Logo, diferente do MD1 e MD2, esses modelos citados foram capazes de apontar as diferenças entre os fatores dos times, em concordância com os dados apresentados do campeonato. Vale destacar que a maioria das estimativas dos fatores obtidas não foram significativas ao nível de 95%, resultando uma grande incerteza nas previsões das partidas.

O MHE conseguiu através das variáveis número de finalizações, escanteios, faltas e cartões trazer mais informações em relação ao número de gols dos times mandantes e visitantes. Através dele pode-se conhecer determinadas características das equipes participantes do campeonato, como os casos citados na Seção 5.7 do Corinthians-SP e Vitória-BA. Além disso, foi o modelo que apresentou o maior número de estimativas significativas dos times.

Com relação ao tempo necessário para obtenção das amostras dos fatores, em média foi necessária 24 horas para a obtenção das amostras, sendo que no caso do modelo ME foi necessário 6 horas e do modelo MHE foi necessário 36 horas.

Para trabalhos futuros deseja-se aplicar os modelos MD2, MDEST1 e MDEST2 com outras edições do campeonato e comparar aos modelos existentes na literatura. Uma vez constatado que os fatores apresentam evolução ao longo das rodadas, pode-se implementar um modelo hierárquico dinâmico (MHD) apresentado abaixo:

$$log(\lambda_i^t) = \Phi_{i1}^t X_{i1}^t + \Phi_{i2}^t X_{i2}^t + \Phi_{i3}^t X_{i3}^t + \Phi_{i4}^t X_{i4}^t, \quad (6.1)$$

$$log(\lambda_j^t) = \Phi_{j1}^t X_{j1}^t + \Phi_{j2}^t X_{j2}^t + \Phi_{j3}^t X_{j3}^t + \Phi_{j4}^t X_{j4}^t. \quad (6.2)$$

Os coeficientes das equipe evoluem no tempo de acordo com as equações de evolução:

$$\Phi_{i1}^t \sim Normal(\Phi_{i1}^{t-1}, \sigma_{\Phi_{i1}}^2), \quad (6.3)$$

$$\Phi_{i2}^t \sim Normal(\Phi_{i2}^{t-1}, \sigma_{\Phi_{i2}}^2), \quad (6.4)$$

$$\Phi_{i3}^t \sim Normal(\Phi_{i3}^{t-1}, \sigma_{\Phi_{i3}}^2), \quad (6.5)$$

$$\Phi_{i4}^t \sim Normal(\Phi_{i4}^{t-1}, \sigma_{\Phi_{i4}}^2). \quad (6.6)$$

Pode-se também considerar outras variáveis, tais como posse de bola, número reais de chances de gols, número de defesas difíceis, número de passes certos, entre outras, visando melhorar a sua capacidade preditiva. Além disso, pode-se considerar em todos os modelos distribuições *a priori* informativas, visando incorporar informações externas de jornalistas esportivos.

Apêndice A

Cadeias do MHE

Nesse apêndice serão apresentadas nas Figuras A1, A2, A3, A4, A5, A6, A7 e A8 as cadeias do MHE referentes aos coeficientes da regressão. Vale destacar que, similar a todos os modelos implementados na presente dissertação, todas as cadeias atingiram a convergência desejada.

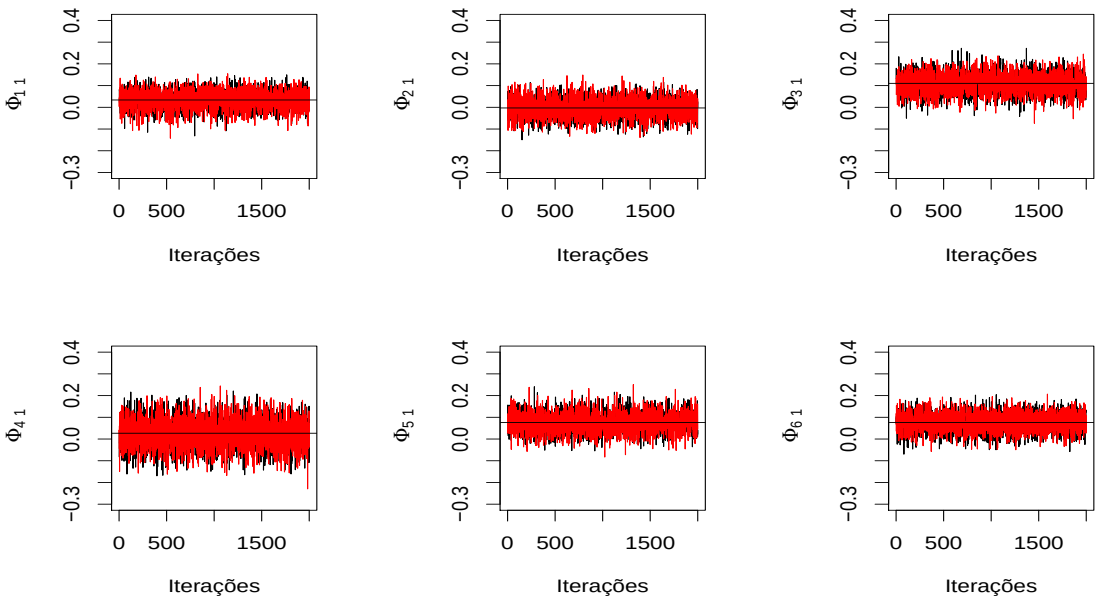


Figura A.1: Coeficientes referentes ao número de finalizações do MHE.

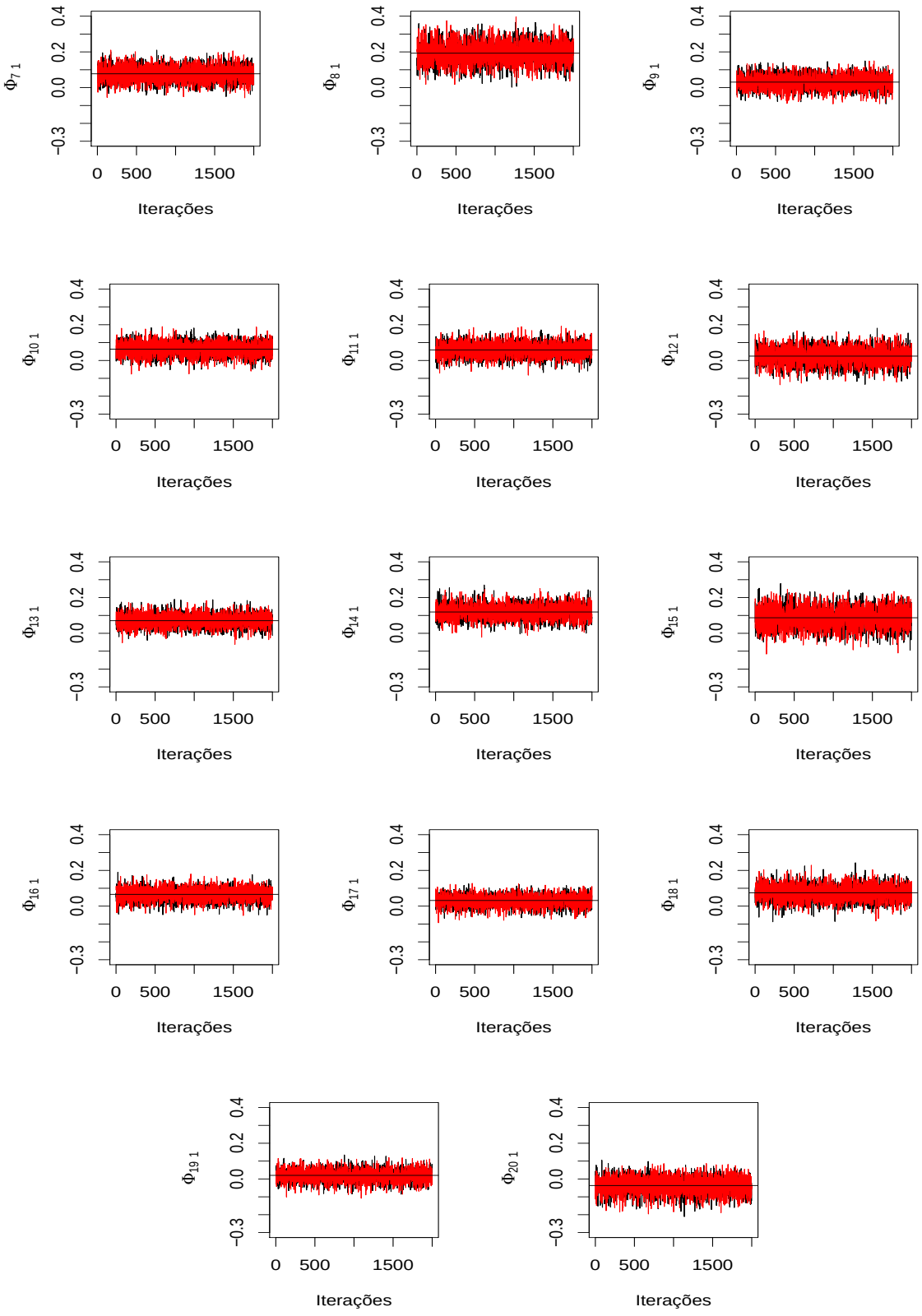


Figura A.2: Coeficientes referentes ao número de finalizações do MHE.

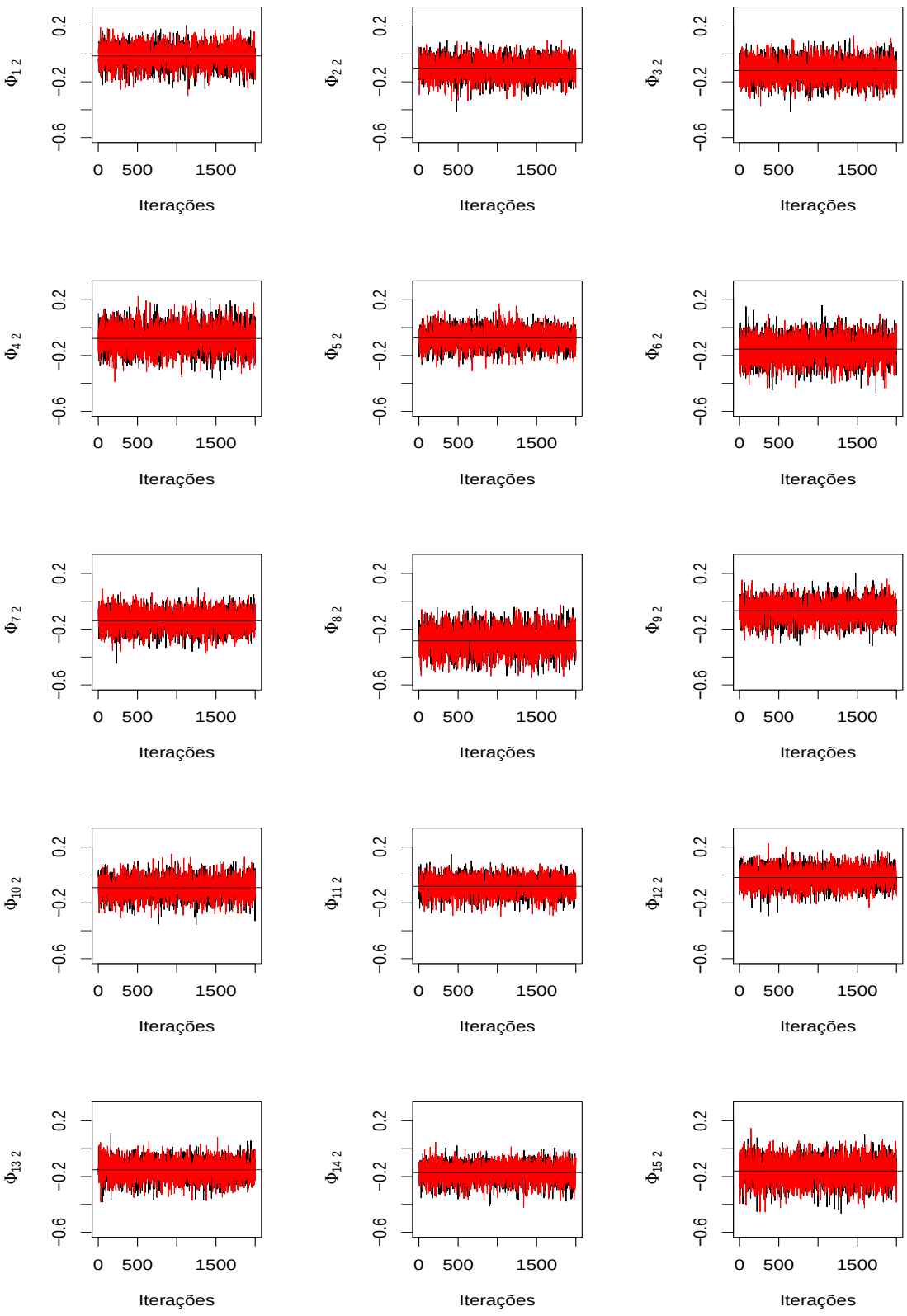


Figura A.3: Coeficientes referentes ao número de escanteios do MHE.

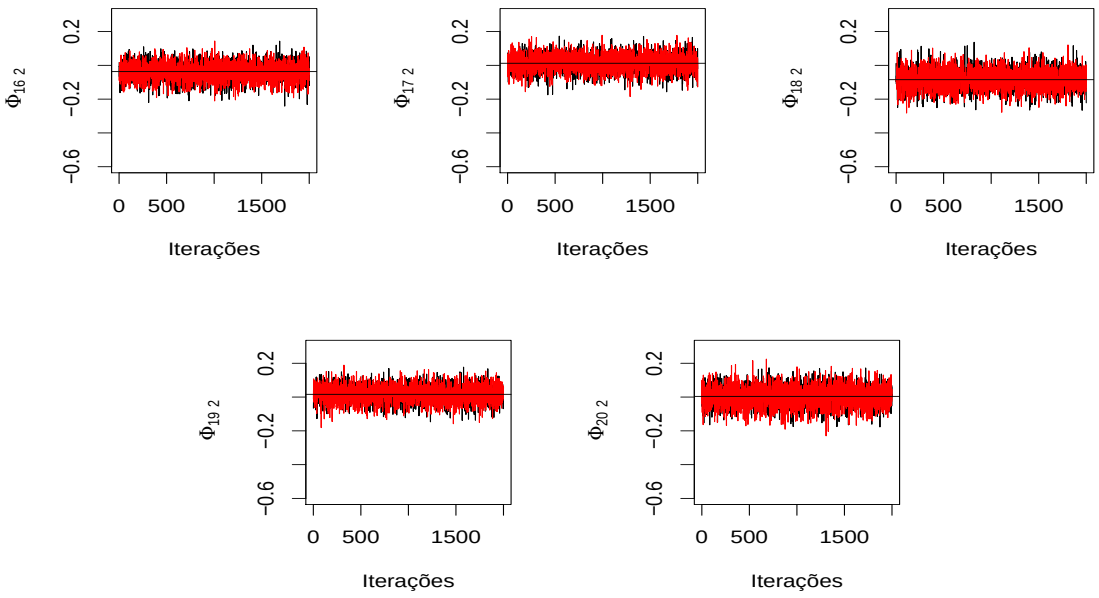


Figura A.4: Coeficientes referentes ao número de escanteios do MHE.

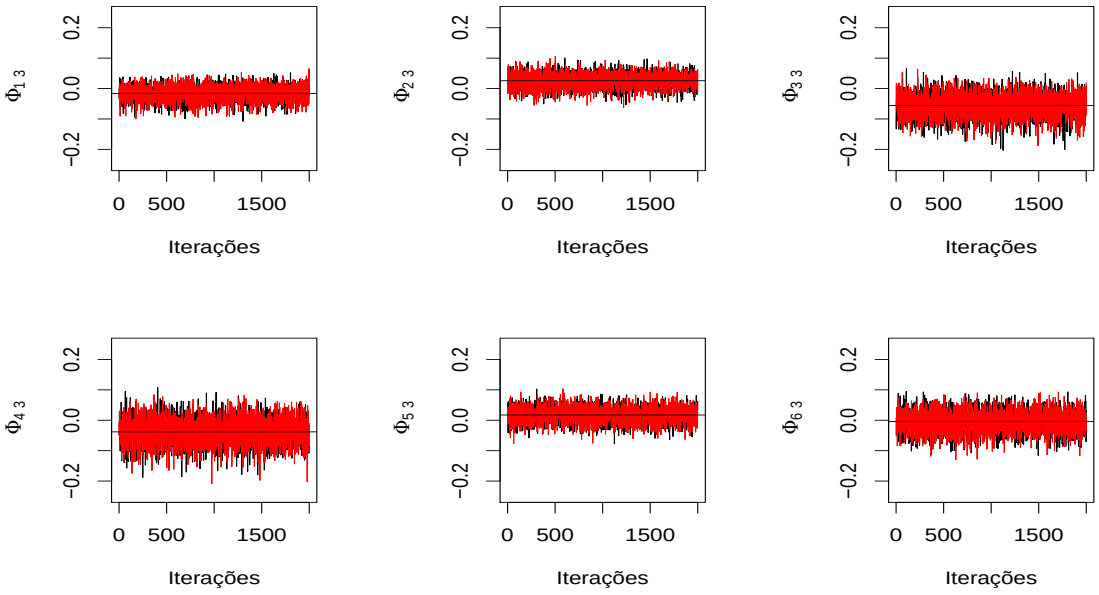


Figura A.5: Coeficientes referentes ao número de faltas do MHE.

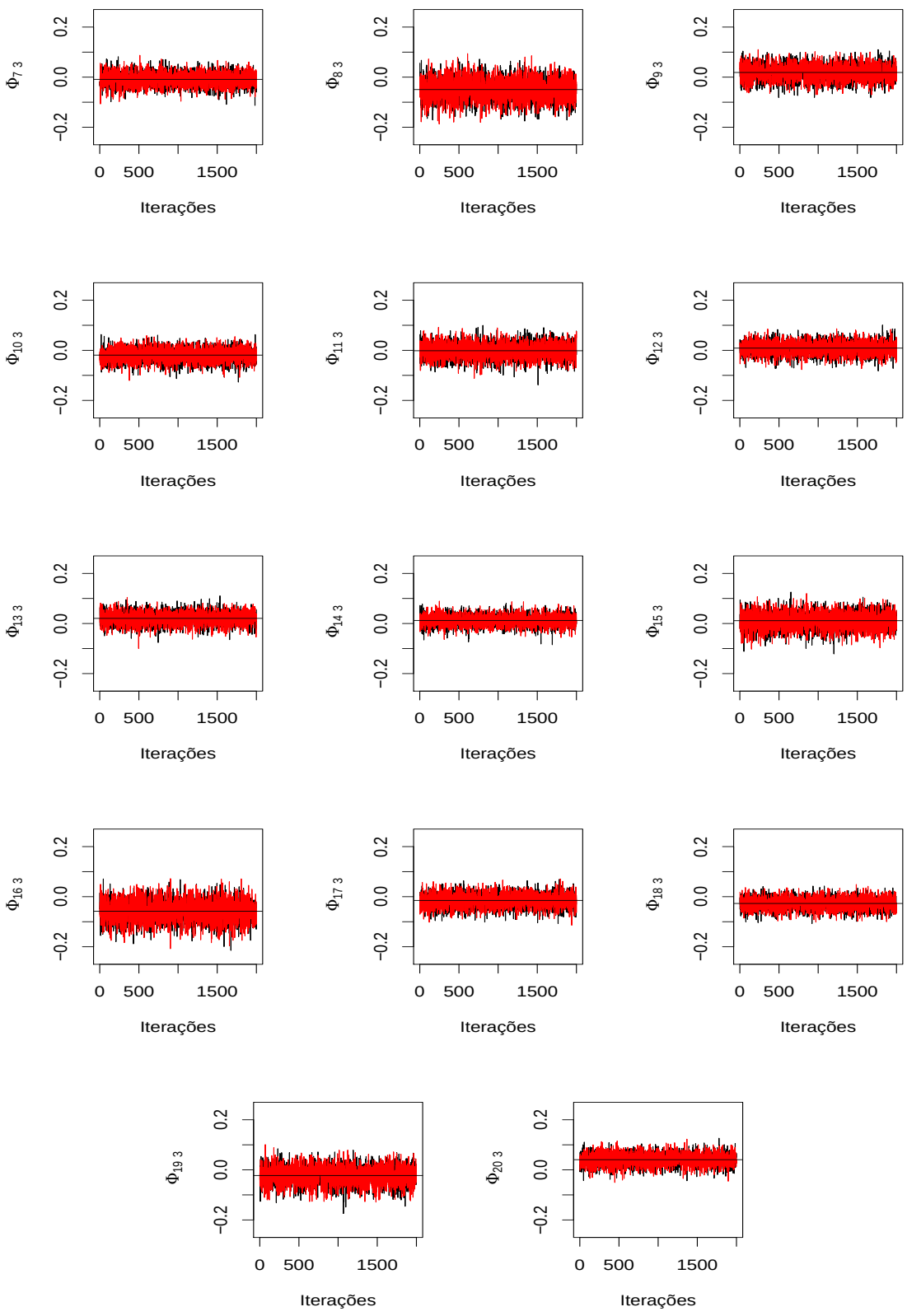


Figura A.6: Coeficientes referentes ao número de faltas do MHE.

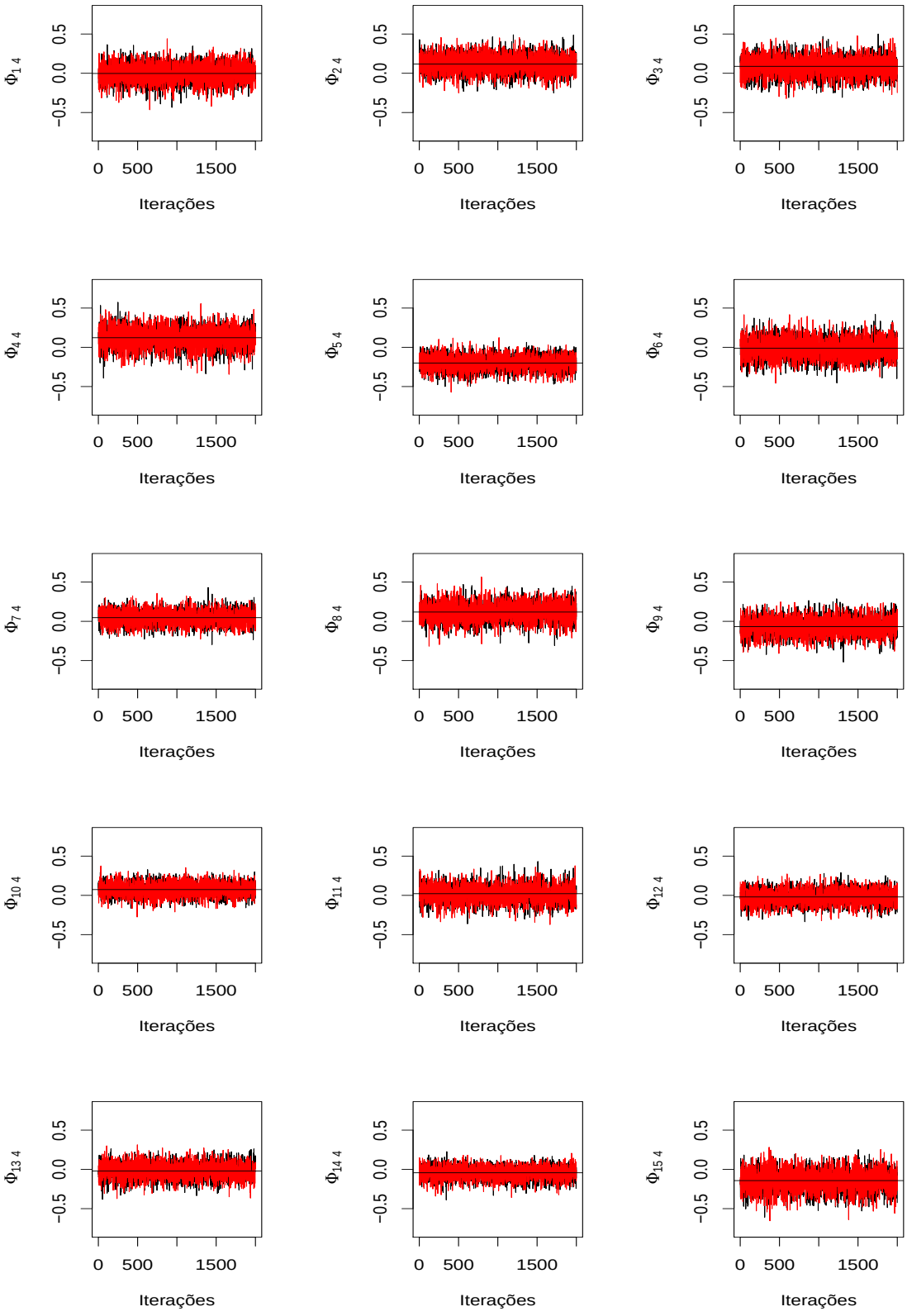


Figura A.7: Coeficientes referentes ao número de cartões do MHE.

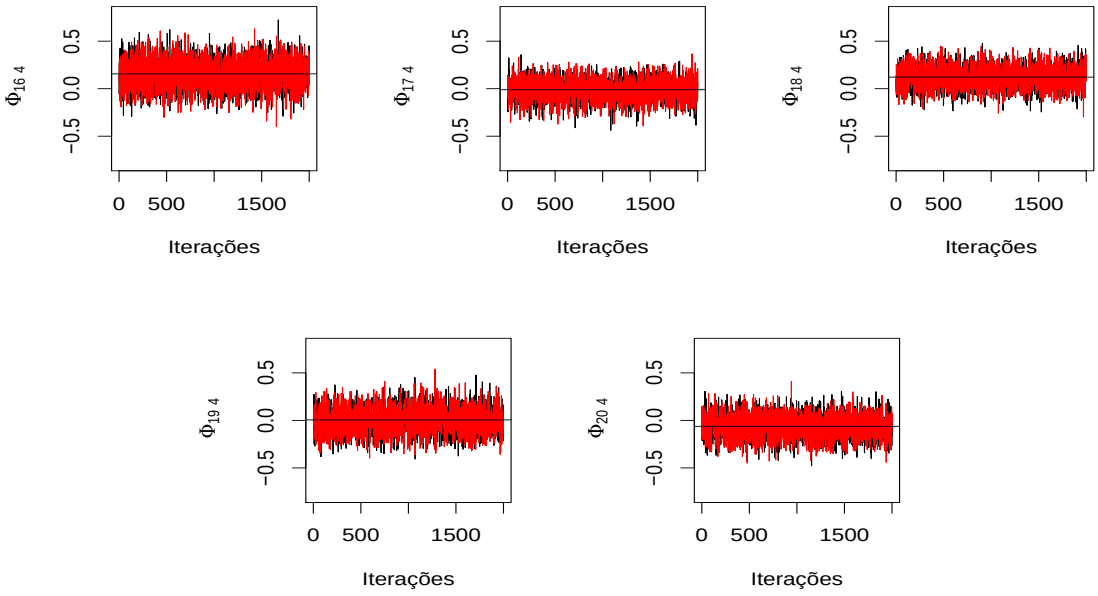


Figura A.8: Coeficientes referentes ao número de cartões do MHE.

Referências Bibliográficas

Casella, G. e Berger, R. (2010) *Inferência Estatística*. São Paulo: Cengage Learning.

CBF (2018) Critérios de desempate do Campeonato Brasileiro de Futebol. URLhttps://cdn.cbf.com.br/content/201703/20170313175547_0.pdf. Data de acesso: 2 jan. 2018.

DeGroot, M. H. e Schervish, M. J. (2012) *Probability and statistics*. Pearson Education, 4nd edn.

Dixon, M. J. e Coles, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series c (Applied Statistics)*, **46**, 265–280.

Dobson, A. J. (2002) *An Introduction to Generalized Linear Models*. New York: Chapman & Hall / CRC, 2nd edn.

Farias, F. (2008) Análise e previsão de resultados de partidas de futebol. Dissertação (Mestrado em Estatística) - Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2008.

Gambeta, W. (2015) *A bola rolou*. São Paulo:SESI.

Gamerman, D. e Lopes, H. F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. New York: Chapman & Hall / CRC.

Gardner, J. (2011) Modeling and simulating football results. URLhttps://www1.maths.leeds.ac.uk/~voss/projects/2010-sports/JamesGardner.pdf.

Geman, S. e Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Globo, E. (2018) Unificação dos títulos Campeonato Brasileiro de Futebol. URLhttp://globoesporte.globo.com/futebol/brasileirao-serie-a/noticia/2010/12/cbf-oficializa-titulos-nacionais-em-cerimonia-com-presenca-de-pele.html. Data de acesso: 1 jan. 2018.

Gneiting, T., Balabdaoui, F. e Raftery, A. E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 243–268.

James, B. (2008) *Probabilidade: um curso em nível intermediário.* IMPA, Rio de Janeiro., 3nd edn.

Knorr-Held, L. (2000) Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **49**, 261–276.

Louzada, F., Suzuki, A., Salasar, L. e ARA, A.; Leite, J. (2015) A simulation-based methodology for predicting football match outcomes considering experts opinions: the 2010 and 2014 football world cup cases. *Pesquisa Operacional*, **35**, 577–598.

Ma, J. e Kockelman, K. (2006) Bayesian multivariate poisson regression for models of injury count, by severity. *Transportation Research Record: Journal of the Transportation Research Board*, 24–34.

Metropolis, N., Rosenbulth, A. W., Rosenbulth, M. N., Teller, A. H. e Teller, E. (1953) Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, **21**, 1087–1091.

Migon, H.S.; Gamerman, D. L. T. (2015) *Statistical Inference. An Integrated Approach.* CRC Press. Taylor& Francis Group, LLC.

Murteira, B. (1990) *Probabilidades e Estatística.* Lisboa:McGraw-Hil.

Nelder, J. A. e Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.

Plummer, M (2013) *JAGS: Just another Gibbs sampler (Version 3.4.0)*. GNU General Public License. URLhttp://mcmc-jags.sourceforge.net.

Poli, G. e Carmona, L. (2009) *Almanaque do futebol Sportv*. Rio de Janeiro: Casa da Palavra: COB Cultural.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/.

Robert, C. e Casella, G. (2004) *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2nd edn.

Rue, H. e Salvesen, O. (2000) Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **49**, 339–418.

Soccerway (2018) Dados coletados. URLhttps://br.soccerway.com/national/brazil/serie-a/2017/regular-season/r39899/. Data de acesso: 5 mai. 2017.

Souza Júnior, O. e Gamerman, D. (2004) Previsão de partidas de futebol usando modelos dinâmicos. *XXXVI Simpósio Brasileiro de Pesquisa Operacional, São João Del Rei*, 649–659.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. e Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society B*, **64**, 583–639.

West, M. e Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag, 2nd edn.

West, M., Harrison, J. e Migon, H. (1985) Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, **80**, 73–83.