Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Equações de Estimação Generalizadas na predição da pontuação de atacantes no Cartola FC

Edvaldo Capobiango Coelho Filho

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Equações de Estimação Generalizadas na predição da pontuação de atacantes no Cartola FC

Edvaldo Capobiango Coelho Filho
Orientador: Prof. Dr. Gustavo Henrique de Araujo Pereira

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
13 de Janeiro de 2021

Edvaldo Capobiango Coelho Filho

Equações de Estimação Generalizadas na predição da pontuação de atacantes no Cartola FC

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Edvaldo Capobiango Coelho Filho e aprovado pela banca examinadora.

São Carlos, 13 de Janeiro de 2021

Banca Examinadora:

- Prof. Dr. Gustavo Henrique de Araujo Pereira (orientador)
- Prof. Dr. Márcio Luis Lanfredi Viola
- Prof. Dr. Afrânio Márcio Corrêa Vieira

Por não medirem esforços para que este momento fosse possível,

dedico este trabalho aos meus pais,

Edvaldo e Niara.

Agradecimentos

Agradeço aos meus pais, Edvaldo e Niara, por serem a minha fonte de inspiração e superação, meus maiores incentivadores, acreditarem na minha capacidade, em momentos que até mesmo eu duvidava, e por todo o amor de sempre, sem eles com certeza eu não teria conseguido chegar até aqui e concluir este ciclo.

Agradeço ao meu cachorro, Rex, por me descontrair mesmo nos piores momentos.

Agradeço ao meu orientador, Gustavo, por todos os ensinamentos, conversas, conselhos, paciência, parceria e ajuda em todos os momentos, sua presença foi crucial.

Agradeço aos demais professores e funcionários do Departamento de Estatística da UFSCar, em especial ao Afrânio e Viola, por terem aceitado fazer parte da banca examinadora do meu trabalho, fazendo ótimas contribuições.

Agradeço aos meus amigos, por estarem comigo em diversos momentos, sejam eles de alegria ou tristeza, aguentarem meus dramas e não desistirem de mim.

Agradeço aos meus familiares, por fazerem parte do que sou hoje.

Agradeço a todos que, de alguma forma, contribuíram para a realização desta etapa.

Resumo

A estatística desempenha um importante papel no meio futebolístico, estando cada vez mais presente, por exemplo, na tomada de decisão dos clubes para a contratação de jogadores. Um outro seguimento que ela está inserida no mundo do futebol é nos *fantasy games*, sendo que o jogo referente à série A do Campeonato Brasileiro é popularmente conhecido por Cartola FC. Coletamos dados pertencentes a alguns atacantes do Cartola FC, através dos *scouts* disponibilizados pelo jogo, com o intuito de predizer a pontuação dos jogadores. Para conseguir tal objetivo utilizaremos os modelos lineares generalizados e as equações de estimação generalizadas, que podem ser consideradas uma extensão da primeira metodologia, mas assumem a existência de uma estrutura de correlação na modelagem dos dados. Essa metodologia é muito usada em estudos longitudinais, em que informações de um mesmo indivíduo são coletadas diversas vezes ao longo do tempo que, no caso deste trabalho, são as rodadas do segundo turno da série A do Campeonato Brasileiro de 2019. Dentre os diversos modelos ajustados, o que apresentou melhor poder preditivo foi um via equação de estimação generalizada, o que é coerente tendo em vista a dependência entre as observações de um mesmo jogador. Porém, alguns modelos lineares generalizados também apresentaram bons desempenhos em relação a predição.

Palavras-chave: *Cartola FC, equações de estimação generalizadas, estruturas de correlação, futebol, modelos lineares generalizados, predição.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O Brasil é mundialmente conhecido como o "país do futebol", seja por ser a seleção que ostenta mais títulos no mundo futebolístico com cinco Copas do Mundo ou por ser um verdadeiro celeiro de craques, como Pelé, que é considerado por muitos como o maior jogador de todos os tempos e foi eleito pela Federação Internacional de Futebol como o Melhor Jogador do Século XX (FIFA, 2000). Nesse contexto, nasce a paixão da maior parte dos brasileiros pelo futebol.

O principal torneio de times no Brasil é a série A do Campeonato Brasileiro, organizado pela Confederação Brasileira de Futebol (CBF). O formato atual é dado pelo sistema de pontos corridos, em dois turnos de 19 partidas, em que cada uma das vinte equipes participantes enfrenta os outros adversários, sendo uma vez em seu estádio e a outra no de seu adversário, totalizando 38 rodadas. Os times recebem três pontos em caso de vitória, um em caso de empate e não são atribuídos pontos para derrotas. São rebaixadas para a série B as quatro equipes com menor quantidade de pontos acumulados após a última rodada. Os seis primeiros colocados participam da Copa Libertadores do ano seguinte, o maior torneio futebolístico da América. É campeão o time que contabilizar o maior número de pontos ao final do campeonato (CBF, 2019).

A estatística desempenha um papel essencial no seguimento de esportes, seja para a contratação de jogadores por parte das equipes a partir de dados dos mesmos, como para escalar equipes virtuais de jogadores reais de um esporte profissional no mundo dos *fantasy games*, que são jogos *online* em que os usuários assumem o papel de "treinador". Essas equipes competem com base no desempenho estatístico dos jogadores escalados nas partidas reais. Este desempenho é convertido em pontos de acordo com regras estabelecidas de cada jogo. No embalo do sucesso dos *fantasy games* da NFL e da NBA, que são,

respectivamente, as ligas profissionais de futebol americano e de basquete dos Estados Unidos, esportes mais populares no país, foi lançado no Brasil em 2005 o Cartola FC.

O Cartola FC, que foi criado pela Rede Globo e promovido pelo canal de TV por assinatura Sportv, é o *fantasy game* referente à série A do Campeonato Brasileiro. Cada usuário começa a temporada com um patrimônio de 100 cartoletas (moeda virtual do jogo, sem valor na vida real). A cada rodada, o usuário opta por uma entre as sete opções de esquema tático oferecidas e escolhe onze jogadores e um técnico para montar seu time dentro do seu limite de cartoletas na rodada em questão. O valor dos jogadores varia de rodada a rodada de acordo com o seu desempenho real, podendo, assim, valorizar ou desvalorizar e, com isso, o patrimônio do usuário é alterado ao longo do campeonato. A lógica do jogo é a de que, por exemplo, se um jogador escalado por você marca gols, isso lhe renderá pontos. Por outro lado, se o seu goleiro escolhido sofre muitos gols, você perde pontos. Ou seja, o sucesso do usuário no jogo vai depender do desempenho real dos atletas escalados no time virtual dentro de campo durante cada rodada do Campeonato Brasileiro. A partir de 2018, o Cartola FC apresentou uma novidade, a opção do Capitão, em que o jogador escolhido para tal cargo no time virtual tem sua pontuação dobrada. Em 2019, o jogo contou com um recorde, cerca de 10 milhões de usuários foram cadastrados.

As estatísticas utilizadas no CartolaFC (2019), chamadas de "*scouts*", e suas respectivas pontuações são:

- RB - Roubada de bolas (1.5);
- G - Gol (8.0);
- A - Assistência (5.0);
- SG - Jogos sem sofrer gols (5.0);
- FS - Falta sofrida (0.5);
- FF - Finalização para fora (0.8);
- FD - Finalização defendida (1.2);
- FT - Finalização na trave (3.0);
- DD - Defesa difícil (3.0);
- DP - Defesa de pênalti (7.0);

- GC - Gol contra (-5.0);
- CV - Cartão vermelho (-5.0);
- CA - Cartão amarelo (-2.0);
- GS - Gol sofrido (-2.0);
- PP - Pênalti perdido (-4.0);
- FC - Falta cometida (-0.5);
- I - Impedimento (-0.5);
- PE - Passe errado (-0.3).

Vale ressaltar que a pontuação atribuída aos técnicos é dada pela média de pontos dos jogadores da equipe do mesmo. Ainda, destaca-se que algumas estatísticas são exclusivas da posição do jogador, como DD, DP e GS, que são de goleiros, e SG, que é para jogadores do sistema defensivo, ou seja, goleiros, laterais e zagueiros. A maioria dos *scouts* são baseados nas súmulas das partidas disponibilizadas pela CBF. Por outro lado, algumas, como DD e RB, são interpretativas, sendo computadas manualmente de acordo com a opinião dos funcionários do Cartola FC (GloboEsporte, 2019).

Para este trabalho será criado um banco de dados real com diversos atacantes, tendo pelo menos um de cada uma das vinte equipes, no qual serão observadas, em cada uma das 19 rodadas do segundo turno, algumas informações referente a cada jogador. O principal objetivo é ajustar modelos de equações de estimação generalizadas com esses dados a fim de predizer a pontuação no Cartola FC dos atacantes no segundo turno da série A do Campeonato Brasileiro de 2019. É esperado que a pontuação do jogador em determinada rodada tenha correlação com a pontuação desse mesmo jogador na rodada anterior, o que justifica o uso das equações de estimação generalizadas. No entanto, como o foco do trabalho está na predição, também serão ajustados modelos lineares generalizados, pois pode ser que o seu poder preditivo seja melhor, apesar de, pensando na teoria, não ser a metodologia mais adequada para esse trabalho, visto que não considera a presença de correlação entre observações de um mesmo indivíduo.

Este trabalho está organizado da seguinte maneira. No Capítulo 2, apresentamos a estrutura e os componentes de um Modelo Linear Generalizado, descrevendo as

funções de ligação mais utilizadas, procedimentos para estimação dos parâmetros pelo método da máxima verossimilhança, testes de hipóteses, método de seleção de variáveis e também uma breve apresentação sobre análise de diagnóstico. Em seguida, no Capítulo 3, abordaremos as principais características das Equações de Estimação Generalizadas, apresentando algumas estruturas da matriz de correlação de trabalho, meios de estimações dos parâmetros, testes de hipóteses e métodos de seleção para a estrutura de correlação e também para as variáveis e uma breve apresentação sobre análise de diagnóstico. O Capítulo 4 consiste na aplicação dos dados, no qual serão descritas as variáveis presentes no banco de dados, realizaremos suas análises descritivas e faremos a modelagem dos dados, ajustando modelos de EEG e MLG, além de verificarmos seus respectivos poderes preditivos. Por fim, no Capítulo 5 é feita a conclusão do trabalho.

Capítulo 2

Modelos Lineares Generalizados

Neste capítulo são apresentadas, na Seção 2.1, a estrutura e os componentes de um modelo linear generalizado. Na Seção 2.2 apresentamos as funções de ligação mais utilizadas. A estimação dos parâmetros do modelo é abordada na Seção 2.3. Na Seção 2.4 discutimos sobre testes de hipóteses. O método de seleção de variáveis *stepwise* via AIC é exposto na Seção 2.5 e uma breve apresentação sobre análise de diagnóstico é mostrada na Seção 2.6.

2.1 Especificação do modelo

Muitos estudos estatísticos têm como principal interesse avaliar a relação funcional entre variáveis, isto é, verificar a influência que uma ou mais variáveis preditoras (X), têm sobre uma variável de interesse, a qual é chamada de variável resposta (Y). Este problema é, em geral, solucionado pelos modelos de regressão.

Durante muitos anos, apesar de já haver outros meios para realizar tal análise, o mais usual era através do Modelo Linear Geral, o qual tem como suposição a normalidade da variável resposta. No entanto, nem sempre a distribuição mais adequada para a variável resposta é a Normal. Dessa forma, Nelder e Wedderburn (1972) unificaram várias classes de modelos, que já existiam, em uma só e, assim, propuseram os Modelos Lineares Generalizados (MLGs), que flexibilizam o uso de outras distribuições para a variável resposta, desde que a mesma pertença à família exponencial linear.

Sejam $Y_1, ..., Y_n$ variáveis aleatórias independentes. Assume-se que a função densidade de probabilidade de Y_i pertence à família exponencial, com $i = 1, 2, ..., n$, que é dada por

$$f(y_i, \theta_i, \phi) = \exp\Big\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\Big\}, \tag{2.1}$$

sendo $E(Y_i) = \mu_i = b'(\theta_i)$ o primeiro momento de Y_i, $Var(Y_i) = \phi^{-1}V_i = \phi^{-1}b''(\theta_i)$ o segundo momento de Y_i, em que $V_i = d\mu_i/d\theta_i$ é denominada função de variância, $\phi^{-1} > 0$ é o parâmetro de dispersão e $c(\cdot)$ é uma função conhecida.

Um MLG é definido por (2.1) acrescido do componente sistemático

$$g(\mu_i) = \eta_i, \tag{2.2}$$

em que $\eta_i = \boldsymbol{x_i}^T\boldsymbol{\beta}$ é o preditor linear, sendo que $\boldsymbol{x_i} = (1, x_{i1}, ..., x_{ip})^T$ é o vetor das variáveis preditoras na observação i, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é um vetor de parâmetros desconhecidos a serem estimados, geralmente por máxima verossimilhança e, $g(\cdot)$ é uma função estritamente monótona e duplamente diferenciável, denominada função de ligação.

Segundo Paula (2004), os três componentes que compõem um MLG são:

- Componente aleatório: representado por um conjunto de variáveis independentes, $Y_1, Y_2, ..., Y_n$, com distribuição pertencente à família exponencial linear com parâmetros ϕ e θ_i, $i = 1, 2, ..., n$;

- Componente sistemático (não aleatório): representado pelo termo $\boldsymbol{x_i}^T\boldsymbol{\beta} = \eta_i$. O componente linear engloba tanto o vetor de parâmetro, como as variáveis preditoras. Assim como no Modelo Linear Geral, o componente sistemático é linear nos parâmetros;

- Função de ligação: relaciona o componente aleatório ao sistemático, sendo $g(\cdot)$ uma função estritamente monótona e duplamente diferenciável.

Então, ajustar um MLG para uma variável resposta envolve escolher uma distribuição que seja adequada para a variável resposta, selecionar, a partir de algum critério, as variáveis preditoras que entrarão no modelo e, por fim, determinar qual a função de ligação mais adequada (Brolo, 2019).

Para o nosso trabalho, como a variável resposta (pontuação do jogador) será contínua, serão consideradas as distribuições normal, gama e gaussiana inversa, as quais aprofundaremos um pouco mais a seguir.

2.1.1 Distribuição normal

Seja Y uma variável aleatória com distribuição normal com média μ e variância σ^2. Então, no formato da família exponencial linear, a função densidade de probabilidade de Y é dada por (Paula, 2004)

$$f(y;\mu;\sigma^2) = \exp\left\{\frac{1}{\sigma^2}\left[\mu y - \frac{\mu^2}{2}\right] - \frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right]\right\}, \tag{2.3}$$

em que $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^{-2}$ e $c(y,\phi) = \frac{1}{2}\left[\log(\frac{\phi}{2\pi})\right] - \frac{\phi y^2}{2}$.

A partir de (2.3), pode-se provar que

- $E(Y) = \mu = \theta$;
- $Var(Y) = \sigma^2 = \frac{1}{\phi}$.

2.1.2 Distribuição gama

Seja Y uma variável aleatória com distribuição gama com parâmetros μ e ϕ. Então, no formato da família exponencial linear, a função densidade de probabilidade de Y é dada por (Paula, 2004)

$$f(y;\mu;\phi) = \exp\left\{\phi\left[-\frac{y}{\mu} - \log(\mu)\right] - \log\Gamma(\phi) + \phi\log(\phi) + (\phi-1)\log y\right\}, \tag{2.4}$$

em que $\theta = -\frac{1}{\mu}$, $b(\theta) = -\log(-\theta)$ e $c(y,\phi) = -\log\Gamma(\phi) + \phi\log(\phi) + (\phi-1)\log y$.

A partir de (2.4), pode-se provar que

- $E(Y) = \mu = -\frac{1}{\theta}$;
- $Var(Y) = \phi^{-1}\mu^2$.

2.1.3 Distribuição gaussiana inversa

Seja Y uma variável aleatória com distribuição gaussiana inversa com parâmetros μ e ϕ. Então, no formato da família exponencial linear, a função densidade de probabilidade de Y é dada por (Paula, 2004)

$$f(y;\mu;\phi) = \exp\left\{\phi\left[-\frac{y}{2\mu^2} + \frac{1}{\mu}\right] - \frac{1}{2}\left[\log\left(\frac{2\pi y^3}{\phi}\right) + \frac{\phi}{y}\right]\right\}, \tag{2.5}$$

em que $\theta = -\frac{1}{\mu^2}$, $b(\theta) = -(-2\theta)^{\frac{1}{2}}$ e $c(y,\phi) = \frac{1}{2}\left[\log\left(\frac{2\pi y^3}{\phi}\right) + \frac{\phi}{y}\right]$.

A partir de (2.5), pode-se provar que

- $E(Y) = \mu = (-2\theta)^{-\frac{1}{2}}$;
- $Var(Y) = \phi^{-1}\mu^3$.

2.2 Função de ligação

A função de ligação $g(\cdot)$ pode ser qualquer função desde que seja estritamente monótona e duplamente diferenciável. Se $g(\mu_i) = \theta_i$, o preditor linear modela o parâmetro canônico θ_i e denominamos esta função de ligação como canônica. O uso de uma função de ligação tem como vantagens a garantia de unicidade do estimador de máxima verossimilhança (EMV) de $\boldsymbol{\beta}$, a simplificação do algoritmo de estimação de $\boldsymbol{\beta}$ e o papel de garantir que μ_i pertença ao suporte da função densidade de probabilidade de Y_i (Pereira, 2019).

A distribuição normal tem como função de ligação canônica a função identidade, que é dada por $g(\mu_i) = \mu_i$. Por outro lado, a distribuição gama tem como função de ligação canônica a função recíproca, que é representada por $g(\mu_i) = \frac{1}{\mu_i}$, e a da gaussiana inversa é a recíproca ao quadrado, dada por $g(\mu_i) = \frac{1}{\mu_i^2}$ (Bové et al., 2011). Porém, no caso dessas duas últimas distribuições, o uso de suas funções de ligações canônicas não são recomendadas por dois motivos. O primeiro é que o seu uso pode produzir valores negativos para $\hat{\mu}_i$, o que não é adequado, visto que, em ambas as distribuições, μ_i é obrigatoriamente positiva, já que $y_i > 0$. Além disso, não é possível interpretar os parâmetros se usamos a função de ligação canônica nesses casos.

Neste contexto, as funções de ligações mais utilizadas para as distribuições gama e gaussiana inversa são a logarítmica e a identidade, as quais abordaremos um pouco mais a fundo. No entanto, vale ressaltar que, apesar de não ser muito usada, uma outra possível função de ligação nesses casos é a raiz quadrada.

2.2.1 Logarítmica

A função de ligação logarítmica é definida por (Dias et al., 2013)

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{x_i}^T\boldsymbol{\beta}, \tag{2.6}$$

e, deste modo, $\mu_i = \exp\left\{\boldsymbol{x_i}^T\boldsymbol{\beta}\right\}$.

Para interpretar os parâmetros do modelo, primeiramente, considere $x_{ij} = l$. Ao fixar-se as outras variáveis, temos que

$$\mu_i = \exp\left\{\beta_0 + \beta_1 x_{i1} + ... + \beta_j l + ... + \beta_p x_{ip}\right\}. \tag{2.7}$$

Em contrapartida, se $x_{ij} = l + 1$,

$$\mu_i = \exp\left\{\beta_0 + \beta_1 x_{i1} + ... + \beta_j (l+1) + ... + \beta_p x_{ip}\right\}. \tag{2.8}$$

Nota-se que o valor de μ_i expresso em (2.8) é (2.7) $\exp\{\beta_j\}$. Então, $\exp\{\beta_j\}$ é o valor pelo qual é multiplicado a média da variável resposta quando aumentamos x_{ij} em uma unidade e mantemos as demais variáveis preditoras constantes.

2.2.2 Identidade

A função de ligação identidade é definida por (Dias et al., 2013)

$$g(\mu_i) = \mu_i = \boldsymbol{x_i}^T\boldsymbol{\beta}. \tag{2.9}$$

Para interpretar os parâmetros do modelo, primeiramente, considere $x_{ij} = l$. Ao fixar-se as outras variáveis, temos que

$$\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_j l + ... + \beta_p x_{ip}. \tag{2.10}$$

Em contrapartida, se $x_{ij} = l + 1$,

$$\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_j (l+1) + ... + \beta_p x_{ip}. \tag{2.11}$$

Nota-se que o valor de μ_i expresso em (2.11) é (2.10) $+ \beta_j$. Então, a interpretação dos parâmetros do modelo é feita da mesma maneira do Modelo Linear Geral, ou seja, o valor de β_j indica a alteração na média da variável resposta quando aumentamos x_{ij} em uma unidade e mantemos as demais variáveis preditoras constantes.

2.3 Estimação dos parâmetros

Os parâmetros do modelo são estimados via método de máxima verossimilhança, atendendo suas propriedades gerais, no qual obtemos inicialmente o logaritmo da função de verossimilhança, dado por (Neter et al., 1996)

$$l(\boldsymbol{\beta};\phi) = \sum_{i=1}^{n}\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\} = \sum_{i=1}^{n} l_i(\boldsymbol{\beta};\phi). \tag{2.12}$$

em que $l_i(\boldsymbol{\beta};\phi)$ é o componente do logaritmo da função de verossimilhança para a i-ésima observação.

2.3.1 Estimação de β

Para obtermos o estimador de máxima verossimilhança (EMV) do parâmetro $\boldsymbol{\beta}$, derivamos $l(\boldsymbol{\beta};\phi)$ em relação a β_r com o intuito de obter $U_{\beta_r}(\boldsymbol{\gamma})$, que é um componente do vetor escore $\boldsymbol{U}_{\beta}(\boldsymbol{\gamma})$, sendo que $\boldsymbol{\gamma} = (\beta_0, \beta_1, ..., \beta_p, \phi)^T$ é o vetor de parâmetros. Então, o componente é representado por (Pereira, 2019)

$$U_{\beta_r}(\boldsymbol{\gamma}) = \frac{\partial l(\boldsymbol{\beta};\phi)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta};\phi)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta};\phi)}{\partial \theta_i}\, \frac{\partial \theta_i}{\partial \mu_i}\, \frac{\partial \mu_i}{\partial \eta_i}\, \frac{\partial \eta_i}{\partial \beta_r}\,, \tag{2.13}$$

em que:

- $\frac{\partial l_i(\boldsymbol{\beta};\phi)}{\partial \theta_i} = \phi y_i - \phi b'(\theta_i) = \phi[y_i - b'(\theta_i)] = \phi[y_i - \mu_i]$;
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)} = \frac{1}{V_i}$;
- $\frac{\partial \mu_i}{\partial \eta_i}$ varia de acordo com a função de ligação $g(\cdot)$;
- $\frac{\partial \eta_i}{\partial \beta_r} = x_{ir}$.

Assim, temos

$$U_{\beta_r}(\boldsymbol{\gamma}) = \phi(y_i - \mu_i)\, \frac{1}{V_i}\, \frac{\partial \mu_i}{\partial \eta_i}\, x_{ir} = \phi(y_i - \mu_i) w_i\, \frac{\partial \eta_i}{\partial \mu_i}\, x_{ir}, \tag{2.14}$$

sendo $w_i = \frac{1}{V_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$.

Matricialmente podemos então escrever o vetor escore referente ao vetor de parâmetros $\boldsymbol{\beta}$ como

$$\boldsymbol{U}_{\beta}(\boldsymbol{\gamma}) = \phi \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{G} (\boldsymbol{y} - \boldsymbol{\mu}), \tag{2.15}$$

em que $\boldsymbol{W} = \text{diag}\ \{w_1, ..., w_n\}$, $\boldsymbol{G} = \text{diag}\ \{g'(\mu_1), ..., g'(\mu_n)\}$e $\boldsymbol{y} - \boldsymbol{\mu} = \begin{bmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{bmatrix}$.

O EMV de $\boldsymbol{\beta}$ é obtido igualando-se $\boldsymbol{U}_{\beta}(\boldsymbol{\gamma})$ a um vetor de zeros e resolvendo-se o sistema de $p+1$ equações. Porém, geralmente esse sistema envolve equações não lineares e, com isso, não possui solução analítica. Assim, uma alternativa para este problema é a utilização de métodos numéricos usando algoritmos iterativos.

Um dos métodos numéricos que possibilitam estimar $\boldsymbol{\beta}$ é o método de Newton-Raphson, que é baseado na expansão de Taylor de uma função genérica $f(\cdot)$ para a solução de $f(x) = 0$ em torno do ponto x_0, isto é,

$$f(x) \cong f(x_0) + (x - x_0) f'(x_0) = 0. \tag{2.16}$$

Desenvolvendo-se a Equação (2.16), obtém-se

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{2.17}$$

Em termos de um processo iterativo, temos

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}. \tag{2.18}$$

De forma análoga, a versão matricial de (2.18) é

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\boldsymbol{J}^{(m)}]^{-1} \boldsymbol{U}^{(m)}, \tag{2.19}$$

em que $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\beta}^{(m+1)}$ são os vetores de parâmetros estimados nos passos m e $m+1$, $\boldsymbol{U}^{(m)}$ o vetor escore no passo m, definido na Equação (2.15) e, por fim, $[\boldsymbol{J}^{(m)}]^{-1}$ é a inversa da matriz de informação observada avaliada no passo m, que é obtida na posição $(r+1) \times (s+1)$ pelo termo $\frac{-\partial^2 l(\boldsymbol{\beta};\phi)}{\partial \beta_r \beta_s}$.

Uma variação mais conveniente do método Newton-Raphson é o método Scoring de Fisher, que utiliza a matriz de informação de Fisher, denotada por K, no lugar da matriz de informação observada. Desse modo, a expressão se dá por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\boldsymbol{K}^{(m)}]^{-1}\boldsymbol{U}^{(m)}, \tag{2.20}$$

em que $\boldsymbol{K}$ é obtida na posição $(r+1) \times (s+1)$ através do termo $-E\left(\frac{\partial^2 l(\boldsymbol{\beta};\phi)}{\partial\beta_r\beta_s}\right)$. Pode-se provar que a informação de Fisher referente ao vetor de parâmetros $\boldsymbol{\beta}$ é dada por (Cordeiro e McCullagh, 1991)

$$\boldsymbol{K}_{\beta\beta}(\boldsymbol{\gamma}) = \phi\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}. \tag{2.21}$$

A Equação (2.20) tem a desvantagem de requerer um valor inicial para $\boldsymbol{\beta}$. Então, para que isso não seja necessário, a reescrevemos da seguinte maneira, a qual dá origem ao método denominado de mínimos quadrados reponderados iterativos (Pedroso, 2007)

$$\boldsymbol{\beta}^{(m+1)} = [\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{(m)}, \tag{2.22}$$

sendo $\boldsymbol{z}^{(m)} = [\boldsymbol{\eta}^{(m)} + \boldsymbol{G}^{(m)}(\boldsymbol{y} - \boldsymbol{\mu}^{(m)})]$.

Como $g(\mu_i) = \boldsymbol{x_i}^T\boldsymbol{\beta} = \eta_i$, um valor inicial para η_i é $g(y_i)$. Dessa forma, com a Expressão (2.22), não precisamos de um valor inicial para $\boldsymbol{\beta}$, e sim para η_i, o qual temos um valor inicial mais intuitivo.

Assim, de acordo com Pereira (2019), os 5 passos para o processo iterativo são

1. Dar um valor inicial para $\boldsymbol{\eta}^{(0)}$ e com isso obter $\boldsymbol{W}^{(0)}$ e $\boldsymbol{z}^{(0)}$;
2. A partir da Expressão (2.22) obter $\boldsymbol{\beta}^{(1)}$;
3. Tendo $\boldsymbol{\beta}^{(1)}$, atualiza-se $\boldsymbol{\eta}^{(1)}$ e $\boldsymbol{\mu}^{(1)}$ e calcula-se $\boldsymbol{W}^{(1)}$ e $\boldsymbol{z}^{(1)}$;
4. A partir da Expressão (2.22) obter $\boldsymbol{\beta}^{(2)}$;
5. Repetir os passos 3. e 4. até a convergência.

Dentre os inúmeros critérios de parada existentes para analisar a convergência, um deles pode ser expresso como

$$\sum_{r=1}^{p}\left(\frac{\beta_r^{(m+1)} - \beta_r^{(m)}}{\beta_r^{(m)}}\right)^2 < \xi, \tag{2.23}$$

sendo ξ um pequeno valor fixado de parada.

2.3.2 Estimação de ϕ

Segundo Demétrio (2001), para as distribuições Poisson e Binomial, ϕ é unitário e, portanto, não precisa ser estimado. Porém, para as demais distribuições da família exponencial, obtém-se o estimador de máxima verossimilhança (EMV) do parâmetro ϕ, derivando $l(\boldsymbol{\beta};\phi)$ em relação a ϕ, tendo assim a função escore para ϕ, a qual é dada por

$$U_\phi(\boldsymbol{\gamma}) = \frac{\partial l(\boldsymbol{\beta};\phi)}{\partial \phi} = \sum_{i=1}^{n}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n}\frac{\partial c(y_i;\phi)}{\partial \phi}. \tag{2.24}$$

O parâmetro ϕ também pode ser estimado pelo método dos momentos, que ao contrário do estimador de máxima verossimilhança, tem forma fechada qualquer que seja a distribuição da variável resposta.

2.4 Testes de hipóteses

Considere $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_a \\ \boldsymbol{\beta}_b\end{bmatrix}$. O interesse é testar $H_0 : \boldsymbol{\beta}_b = \boldsymbol{\beta}_b^{(0)} \times H_1 : \boldsymbol{\beta}_b \neq \boldsymbol{\beta}_b^{(0)}$, em que $\boldsymbol{\beta}_b^{(0)}$ é um vetor de constantes conhecidas. Esse caso engloba tanto o teste para um único parâmetro como para vários, visto que $\boldsymbol{\beta}_b^{(0)}$ pode ser tanto um vetor com apenas um escalar quanto um vetor de vários escalares.

O teste para vários parâmetros é muito comum quando temos variáveis qualitativas e queremos testar todos os parâmetros relacionados as *dummies* de uma variável qualitativa. Nesse caso, testamos se a média da variável resposta varia entre os níveis de uma variável qualitativa.

Dentre as estatísticas existentes para testar essa hipótese, estão a do teste da razão de verossimilhança e a do teste de Wald. Ambas têm, sob H_0, distribuição assintótica qui-quadrado com K_b graus de liberdade, em que K_b é a dimensão do vetor $\boldsymbol{\beta}_b$.

2.4.1 Teste de Wald

De acordo com Turkman e Silva (2000), quando é de interesse testar um único parâmetro β_i, em geral, o teste de Wald é o mais usado. Então, considerando o teste para apenas um parâmetro, a sua estatística teste pode ser representada por

$$Q_W = \frac{(\hat{\beta}_b - \beta_b^{(0)})^2}{\hat{V}ar(\hat{\beta}_b)}\,, \tag{2.25}$$

em que $\hat{\beta}_b$ é o EMV do parâmetro β_b sob H_0 e $\hat{V}ar(\hat{\beta}_b)$ é a $Var(\hat{\beta}_b)$ avaliada em $\boldsymbol{\beta} = \begin{bmatrix} \hat{\beta}_a \\ \hat{\beta}_b \end{bmatrix}$. Vale ressaltar que $\hat{V}ar(\hat{\beta}_b)$ é a inversa de um termo da informação de Fisher e é igual ao estimador do erro padrão de $\hat{\beta}_b^{(0)}$, que é disponibilizado na saída dos softwares.

Rejeita-se H_0 se o valor observado da estatística for superior a $\chi^2_{1,1-\alpha}$, sendo α o nível de significância e $\chi^2_{1,1-\alpha}$ o quantil (1 - α) da χ^2_1.

2.4.2 Teste da razão de verossimilhança

Segundo Turkman e Silva (2000), o teste da razão de verossimilhança, geralmente, é o preferido quando se deseja testar vários $\beta's$ simultaneamente ou quando o intuito é comparar modelos que estão encaixados, ou seja, modelos em que um é submodelo do outro. A sua estatística teste é dada por (Demétrio, 2001)

$$Q_{RV} = 2[l(\hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b, \hat{\phi}) - l(\widetilde{\boldsymbol{\beta}}_a, \widetilde{\boldsymbol{\beta}}_b, \widetilde{\phi})], \tag{2.26}$$

em que $\hat{\boldsymbol{\beta}}_a$, $\hat{\boldsymbol{\beta}}_b$ e $\hat{\phi}$ são os EMV dos parâmetros sob $H_0 \cup H_1$, e $\widetilde{\boldsymbol{\beta}}_a$, $\widetilde{\boldsymbol{\beta}}_b$ e $\widetilde{\phi}$ são os EMV dos parâmetros sob H_0.

Rejeita-se H_0 se o valor observado da estatística for superior a $\chi^2_{K_b,1-\alpha}$, sendo α o nível de significância e $\chi^2_{K_b,1-\alpha}$ o quantil (1 - α) da $\chi^2_{K_b}$.

2.5 Seleção de variáveis

Burnham e Anderson (2004) enfatizam a importância de selecionar modelos baseando-se em princípios científicos. Qualquer procedimento para a seleção de variáveis de um modelo envolve a escolha do melhor ajuste para explicar a variável de interesse considerando todas as combinações possíveis. Dentre os diversos métodos de seleção automática está o *stepwise* via AIC.

O AIC foi proposto por Akaike (1974) e tem como finalidade selecionar modelos utilizando o logaritmo da função de verossimilhança e um termo de penalização baseado na quantidade de parâmetros do mesmo. Assim, escolhe-se o modelo que apresenta, dentre os candidatos, o menor AIC. Deste modo, temos

$$AIC = -2l(\hat{\boldsymbol{\gamma}}) + 2p, \tag{2.27}$$

em que $l(\hat{\boldsymbol{\gamma}})$ é o logaritmo da função de verossimilhança aplicado no EMV de $\boldsymbol{\gamma}$ do modelo, sendo $\boldsymbol{\gamma}$ o vetor de parâmetros do modelo, p é o número de parâmetros do modelo e $2p$ é chamado de termo de penalização, cujo objetivo é permitir a comparação justa de modelos que tenham o número de parâmetros diferentes (Dal Bello, 2010).

O método de seleção de modelos *stepwise* via AIC pode ser resumido em, primeiramente, ajustar todos os modelos possíveis considerando apenas uma variável preditora presente e selecionar aquele que apresenta o menor valor de AIC. Em seguida, ajusta-se todos os modelos que contenham a variável inclusa no primeiro passo mais outra que ainda não foi inserida e, se algum desses apresentar o AIC menor do que o apresentado no primeiro modelo selecionado, a nova variável será inserida. O próximo passo consiste em ajustar todos os modelos que contenham as duas variáveis selecionadas no passo anterior mais outra que ainda não foi inserida e, se algum desses apresentar o AIC menor do que o apresentado no modelo selecionado anteriormente, a nova variável será adicionada. Ainda nesse passo, verifica-se a possibilidade de retirar alguma das variáveis presentes no modelo, dado que com a inclusão da última variável a exclusão de outra pode diminuir o AIC. Dessa maneira, repete-se esses passos de avaliar a inclusão e exclusão das variáveis até que nem a inclusão de alguma variável não presente no modelo e nem a exclusão de alguma variável presente reduza o AIC do modelo.

2.6 Análise de diagnóstico

A análise de diagnóstico é uma importante etapa no ajuste de modelos de regressão. No caso dos MLGs, os elementos da diagonal principal da matriz de projeção (matriz chapéu), os resíduos (sendo os mais utilizados os de Pearson, deviance e quantílico) e a distância de Cook são algumas medidas úteis para detectar pontos alavancas, discrepantes e influentes, respectivamente. No que se diz respeito à avaliação da distribuição escolhida para a variável resposta, usualmente se faz com gráficos de probabilidade normal para os resíduos com a inclusão de uma espécie de "banda de confiança", denominada de envelope simulado.

Como o objetivo do trabalhado é preditivo, comumente não é feita análise de diagnóstico nesse caso, visto que o modelo escolhido será o que apresentar o melhor poder de predição. Sendo assim, o leitor interessado em mais detalhes sobre análise de diagnóstico em MLGs pode consultar em Paula (2004).

Capítulo 3

Equações de Estimação Generalizadas

Neste capítulo são apresentadas, na Seção 3.1, a estrutura e os componentes de um modelo de equação de estimação generalizada. Na Seção 3.2 apresentamos as estruturas da matriz de correlação de trabalho. Os métodos de estimação dos parâmetros do modelo são abordados na Seção 3.3. Na Seção 3.4 discutimos sobre testes de hipóteses. Na Seção 3.5 são vistos métodos de seleção de variáveis. Os métodos de seleções da estrutura de correlação são discutidos na Seção 3.6 e uma breve apresentação sobre análise de diagnóstico é exposta na Seção 3.7.

3.1 Especificação do modelo

Uma suposição muito corrente na análise de modelos de regressão é a de independência entre as observações. No entanto, há situações em que essa suposição pode não fazer muito sentido, como no caso de dados longitudinais, no qual as medidas de um indivíduo são avaliadas repetidas vezes ao longo do tempo e, com isso, é viável considerar a existência de correlação. Diante disso, um dos métodos estatísticos mais utilizados para a resolução de problemas com medidas repetidas foi proposto por Liang e Zeger (1986), as Equações de Estimação Generalizadas (EEGs), que podem ser consideradas uma extensão dos Modelos Lineares Generalizados (MLGs). Apesar de não ser o caso que abordaremos, as EEGs também podem ser usadas quando, ao invés de termos n indivíduos observados em diversos instantes de tempo, temos grupos de indivíduos correlacionados, os quais denominamos de dados agrupados, como, por exemplo, membros de uma mesma família

ou estudantes de uma mesma escola.

Dentre as semelhanças das duas metodologias citadas está o fato de que a distribuição da variável resposta pode pertencer à família exponencial linear nos EEGs, condição obrigatória nos MLGs segundo formulação apropriada originalmente por Nelder e Wedderburn (1972). Ainda, a média e a variância são definidas igualmente para cada observação. Ou seja, no caso dos dados serem longitudinais, considere $\boldsymbol{y_i} = (y_{i1}, ..., y_{in_i})^T$, $i = 1, 2, ..., n$, um vetor $n_i \times 1$ de respostas para cada indivíduo i, e $\boldsymbol{x_{ij}}$ um vetor $p \times 1$ de covariáveis associadas à j-ésima observação do i-ésimo indivíduo, y_{ij}, $j = 1, 2, ..., n_i$. Note que o valor de n_i pode variar de indivíduo para indivíduo. Neste contexto, de maneira análoga aos MLGs, considerando que a distribuição da variável resposta pertence à família exponencial linear, a função densidade de probabilidade de Y_{ij} é dada por

$$f(y_{ij}, \theta_{ij}, \phi) = \exp \phi[y_{ij}\theta_{ij} - b(\theta_{ij})] + c(y_{ij}, \phi), \tag{3.1}$$

em que $E(Y_{ij}) = \mu_{ij} = b'(\theta_{ij})$, $Var(Y_{ij}) = \phi^{-1}V_{ij} = \phi^{-1}b''(\theta_{ij})$ e $V_{ij} = d\mu_{ij}/d\theta_{ij}$ são, respectivamente, a média, a variância e a função de variância do i-ésimo indivíduo em sua j-ésima observação. $\phi^{-1} > 0$ é o parâmetro de dispersão e $c(\cdot)$ é uma função conhecida.

Para definir o modelo, acresce-se em (3.1) o componente sistemático expresso por

$$g(\mu_{ij}) = \eta_{ij}, \tag{3.2}$$

em que $\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é um vetor de parâmetros desconhecidos a serem estimados, neste caso via equacões de estimação generalizadas e, $g(\cdot)$ é uma função estritamente monótona e duplamente diferenciável, denominada função de ligação.

Como aqui, no contexto das EEGs, as observações não são consideradas independentes, temos que, de alguma forma, considerar a presença de uma correlação na modelagem dos dados. Dessa maneira, Liang e Zeger (1986) propuseram a segmentação da matriz de variância e covariância em quatro termos, em que um deles é justamente para especificar a estrutura de correlação dos dados. Sendo assim, dentre as particularidades das EEGs, está a ideia da matriz de correlação de trabalho, $\boldsymbol{R}_i(\boldsymbol{\alpha})$, matriz simétrica $n_i \times n_i$ positiva definida que especifica a estrutura de correlação entre observações de um mesmo indivíduo. Ela é denominada matriz de correlação de trabalho, porque os estimadores dos parâmetros $\boldsymbol{\beta}$ do modelo são consistentes mesmo que a matriz seja especificada

de forma incorreta. Dessa forma, a matriz de variância e covariância da variável resposta para o indivíduo i pode ser decomposta como (Liang e Zeger, 1986)

$$\boldsymbol{\Omega}_i = \phi^{-1}\boldsymbol{A}_i{}^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i{}^{1/2}, \tag{3.3}$$

em que $\boldsymbol{\alpha}$ é um vetor de parâmetros que caracteriza completamente $\boldsymbol{R}_i(\boldsymbol{\alpha})$ e $\boldsymbol{A}_i = \text{diag}\ \{V_{i1}, ..., V_{in_i}\} = \text{diag}\ \{b''(\theta_{i1}), ..., b''(\theta_{in_i})\}$ é uma matriz diagonal com dimensão $n_i \times n_i$.

Vale ressaltar que apesar das observações de um mesmo indivíduo serem correlacionadas, supõe-se que haja independência entre as observações de indivíduos diferentes (Ziegler, 2011).

Como o número de observações pode diferir de indivíduo para indivíduo, o mesmo pode ocorrer com a matriz de correlação. Entretanto, pode-se assumir que $\boldsymbol{R}_i(\boldsymbol{\alpha})$ é completamente especificado pelo vetor de parâmetros $\boldsymbol{\alpha}$, que é o mesmo para todos os indivíduos. Com isso, considera-se $\boldsymbol{R}(\boldsymbol{\alpha})$ para denotar a matriz de correlação de trabalho de todos os indivíduos (Agranonik, 2009). Cada posição de $\boldsymbol{R}(\boldsymbol{\alpha})$ contém valores que estão no intervalo [-1;1], já que correspondem à correlações entre observações de um mesmo indivíduo em diferentes instantes de tempo.

3.2 Estruturas da matriz de correlação de trabalho

Dentre as diversas estruturas da matriz de correlação de trabalho estão a independente, a permutável, a auto-regressiva de primeira ordem e a não estruturada, as quais abordaremos com mais detalhes a seguir. No entanto, apesar de não ser o foco do trabalho, é importante dizer que existem outras possíveis estruturas de correlação, como a estacionária, a não estacionária e a fixa (Hardin e Hilbe, 2013). Para todas as estruturas de matriz de correlação de trabalho, a diagonal principal tem todos seus elementos iguais a 1, porque nessas posições estão sendo avaliadas correlações envolvendo cada observação e ela mesma.

3.2.1 Independente

A estrutura de correlação independente considera a ausência de correlação entre as observações. Ou seja, ela é usada quando assumimos que todas as observações são independentes, como no caso dos MLGs. Sendo assim, essa forma de correlação é expressa

por (Ziegler, 2011)

$$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & se \quad j = j' \\ 0, & se \quad j \neq j' \end{cases} \tag{3.4}$$

Nesse contexto, a estrutura da matriz de correlação é igual a matriz identidade, logo, é dada por

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3.5}$$

3.2.2 Permutável

A estrutura de correlação permutável, também denominada como uniforme ou simetria composta, considera que a correlação entre todas as observações de um mesmo indivíduo i é a mesma. Ela é muito usada quando consideramos amostras por conglomerados, como em estudos domiciliares e familiares. Sendo assim, essa forma de correlação é expressa por (Ziegler, 2011)

$$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & se \quad j = j' \\ \alpha, & se \quad j \neq j' \end{cases} \tag{3.6}$$

Nesse contexto, a estrutura da matriz de correlação de trabalho, a qual todas as correlações de diferentes observações são equivalentes, é dada por (Hardin e Hilbe, 2013)

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix}. \tag{3.7}$$

3.2.3 Auto-regressiva de primeira ordem

A estrutura de correlação auto-regressiva de primeira ordem, ou simplesmente AR(1), considera que a correlação entre as observações de um mesmo indivíduo i diminui

exponencialmente ao longo do tempo. Sendo assim, essa forma de correlação é expressa por (Ziegler, 2011)

$$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & se \quad j = j' \\ \alpha^{|j-j'|}, & se \quad j \neq j' \end{cases} \tag{3.8}$$

Nesse contexto, é compreensível o fato dessa estrutura da matriz de correlação de trabalho ser muito usada para dados longitudinais, visto que ela considera um decréscimo do valor de α conforme aumenta a distância entre j e j' na matriz. Dito isto, o formato da matriz de correlação de trabalho auto-regressiva é dado por (Hardin e Hilbe, 2013)

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{|1-n_i|} \\ \alpha & 1 & \alpha & \cdots & \alpha^{|2-n_i|} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{|3-n_i|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{|n_i-1|} & \alpha^{|n_i-2|} & \alpha^{|n_i-3|} & \cdots & 1 \end{bmatrix}. \tag{3.9}$$

3.2.4 Não estruturada

A estrutura de correlação não estruturada recebe esse nome justamente por não assumir uma estrutura específica e, por conta disso, é considerada a mais geral das estruturas. Sendo assim, essa forma de correlação é expressa por (Ziegler, 2011)

$$Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1, & se \quad j = j' \\ \alpha_{jj'} = \alpha_{j'j}, & se \quad j \neq j' \end{cases} \tag{3.10}$$

Nesse contexto, considerando essa estrutura para a matriz de correlação de trabalho, não é garantido que ela seja invertível e, devido a isso, podem ocorrer problemas numéricos na estimação de seus parâmetros (Hardin e Hilbe, 2013). Dito isto, o formato da matriz de correlação de trabalho não estruturada é dado por

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n_i} \\ \alpha_{12} & 1 & \alpha_{23} & \cdots & \alpha_{2n_i} \\ \alpha_{13} & \alpha_{23} & 1 & \cdots & \alpha_{3n_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{1n_i} & \alpha_{2n_i} & \alpha_{3n_i} & \cdots & 1 \end{bmatrix}. \tag{3.11}$$

3.3 Estimação dos parâmetros

Ao longo desta Seção serão discutidos métodos de estimação dos parâmetros. Na Subseção 3.3.1 serão abordados os métodos de estimação para $\boldsymbol{\beta}$. A seguir, na Subseção 3.3.2, apresentamos as estimações de $Var(\hat{\boldsymbol{\beta}})$. A estimação de ϕ se dará na Subseção 3.3.3. Depois, na Subseção 3.3.4, serão mostradas as estimações de $\boldsymbol{\alpha}$. Por fim, será exposto na Subseção 3.3.5 um algoritmo do processo iterativo para estimação dos parâmetros.

3.3.1 Estimação de $\boldsymbol{\beta}$

A função de quase-verossimilhança, inicialmente proposta por Wedderburn (1974) e posteriormente aperfeiçoada por McCullagh (1983), ao contrário da função de verossimilhança que necessita de uma especificação da distribuição da variável resposta, requer poucas suposições sobre a distribuição da variável resposta. De acordo com Baia (1997), para definir uma função de quase-verossimilhança é necessário apenas de uma relação entre a média e a variância. Sendo assim, supondo que $Y_1, ..., Y_n$ são variáveis aleatórias independentes, com $i = 1, ..., n$, o logaritmo da função de quase-verossimilhança é dado por (Vieira, 2004)

$$Q(\boldsymbol{y}; \boldsymbol{\mu}) = \sum_{i=1}^{n} Q_i(y_i; \mu_i) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi V_i} \, d\mu_i, \tag{3.12}$$

em que μ_i e V_i são, respectivamente, a média e a função de variância de Y_i.

Nesse contexto, a função quase-escore de β é representada por (Lara et al., 2012)

$$\boldsymbol{U}^*_{\beta} = \frac{\partial Q(\boldsymbol{y}; \boldsymbol{\mu})}{\partial \boldsymbol{\beta}} = \phi^{-1} \boldsymbol{D}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \tag{3.13}$$

em que $\boldsymbol{V} = \text{diag}\,\{V_1, ..., V_n\}$, $\boldsymbol{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \boldsymbol{W}^{1/2} \, \boldsymbol{V}^{1/2} \, \boldsymbol{X}$, onde $\boldsymbol{W}$, $\boldsymbol{X}$ e ($\boldsymbol{y}$ - $\boldsymbol{\mu}$) são definidos como nos MLGs.

Analogamente, como na estimação por máxima verossimilhança, para a obtenção da estimativa de $\boldsymbol{\beta}$ via quase-verossimilhança é necessária a solução de $\boldsymbol{U}^*_{\beta} = 0$, que é resolvida por algum método númerico como, por exemplo, o método Scoring de Fisher (Paula, 2004).

Por outro lado, considerando que os dados são longitudinais, um estimador via equações de estimação generalizadas para o vetor de parâmetros $\boldsymbol{\beta}$ é obtido a partir da solução do seguinte sistema de equações (Paula, 2004)

$$\boldsymbol{\Psi}_\beta = \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \tag{3.14}$$

em que, nesse caso, $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} = \boldsymbol{W}_i^{1/2}\boldsymbol{V}_i^{1/2}\boldsymbol{X}_i$, sendo que $\boldsymbol{X}_i$ é uma matriz n_i x p de linhas $\boldsymbol{x}_{ij}^T$, $\boldsymbol{W}_i = \text{diag}\ \{w_{i1}, ..., w_{in_i}\}$ é a matriz de pesos com $w_{ij} = \frac{1}{V_{ij}}\left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2$, $\boldsymbol{V}_i = \text{diag}\ \{V_{i1}, ..., V_{in_i}\}$ e $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in_i})^T$.

Pode-se perceber que a estimação por equações de estimação generalizadas para $\boldsymbol{\beta}$ é uma extensão da estimação do vetor de parâmetros via quase-verossimilhança para dados independentes, visto que ambas são semelhantes, exceto pelo fato de que, para obtê-la a partir da Equação (3.14), tem-se adicionalmente o parâmetro α, devido ao termo $\boldsymbol{\Omega}_i^{-1}$, que incorpora aos dados uma estrutura de correlação.

A solução das equações de estimação generalizadas para a estimação de $\boldsymbol{\beta}$ também é feita a partir de um processo iterativo, sendo comum o uso de uma modificação do método Scoring de Fisher que pode ser representada por (Paula, 2004)

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left\{\sum_{i=1}^{n} \boldsymbol{D}_i^{(m)T}\boldsymbol{\Omega}_i^{-(m)}\boldsymbol{D}_i^{(m)}\right\}^{-1}\left\{\sum_{i=1}^{n} \boldsymbol{D}_i^{(m)T}\boldsymbol{\Omega}_i^{-(m)}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^{(m)})\right\}, \tag{3.15}$$

em que (m) e $(m+1)$ são os passos do processo iterativo, sendo $m = 0, 1, 2, ...$

3.3.2 Estimação de $Var(\hat{\boldsymbol{\beta}})$

Segundo Agranonik (2009), há dois possíveis estimadores para a variância de $\hat{\boldsymbol{\beta}}$. O primeiro, definido como estimador baseado no modelo, que assume que a estrutura da matriz de correlação de trabalho é definida corretamente é dado para o i-ésimo indivíduo por (Paula, 2004)

$$\widehat{\boldsymbol{V}}_{\boldsymbol{M}} = \sum_{i=1}^{n} \hat{\boldsymbol{D}}_i{}^T \hat{\boldsymbol{\Omega}}_i{}^{-1} \hat{\boldsymbol{D}}_i. \tag{3.16}$$

Entretanto, muitas vezes, não se tem segurança que a estrutura da matriz de correlação de trabalho é a correta. Dessa maneira, utiliza-se o estimador robusto (sanduíche), definido por (Paula, 2004)

$$\widehat{\boldsymbol{V}}_{\boldsymbol{R}} = \left(\sum_{i=1}^{n} \hat{\boldsymbol{D}}_i{}^T \hat{\boldsymbol{\Omega}}_i{}^{-1} \hat{\boldsymbol{D}}_i\right)^{-1}\left\{\sum_{i=1}^{n} \hat{\boldsymbol{D}}_i{}^T \hat{\boldsymbol{\Omega}}_i{}^{-1} \boldsymbol{C}_i \hat{\boldsymbol{\Omega}}_i{}^{-1} \hat{\boldsymbol{D}}_i\right\}\left(\sum_{i=1}^{n} \hat{\boldsymbol{D}}_i{}^T \hat{\boldsymbol{\Omega}}_i{}^{-1} \hat{\boldsymbol{D}}_i\right)^{-1}, \tag{3.17}$$

em que $\boldsymbol{C}_i = (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T$ são os resíduos empíricos.

Como esse estimador é assintótico, suas propriedades são melhores satisfeitas quando o número de indivíduos é grande, já que com isso o estimador apresente viés pequeno, mesmo que a matriz de correlação de trabalho esteja incorreta (Rotnitzky e Jewell, 1990). Por outro lado, mesmo se a matriz de correlação de trabalho estiver errada, quando a quantidade de indivíduos for pequena, isto é, n menor do que 20, é mais aconselhável utilizar o estimador baseado no modelo (Agranonik, 2009).

3.3.3 Estimação de ϕ

Liang e Zeger (1986) utilizam o método dos momentos para estimar o parâmetro de dispersão ϕ e os escrevem em função dos resíduos de Pearson. Sendo assim, para a observação y_{ij}, o resíduo de Pearson pode ser expresso por

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{V}_{ij}}}\,, \tag{3.18}$$

em que V_{ij} é a função de variância referente à observação y_{ij} e pode ser visualizada no j-ésimo elemento da diagonal principal de $\boldsymbol{A}_i$.

É importante dizer que usualmente a fórmula dos resíduos de Pearson considera no denominador a variância ao invés da função de variância. No entanto, segundo Venezuela (2003) e Oesselmann (2016) podemos denominar o resíduo utilizado em EEG, apresentado na Equação 3.18, também como resíduo de Pearson.

A partir do resíduo de Pearson, temos que a estimativa de ϕ é dada por (Oesselmann, 2016)

$$\hat{\phi} = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{(r_{ij})^2}{(N-p)} \right\}^{-1}, \tag{3.19}$$

com $N = \sum_{i=1}^{n} n_i$.

3.3.4 Estimação de α

Assim como para a estimação de ϕ, Liang e Zeger (1986) utilizam o método dos momentos para estimar os parâmetros de correlação $\boldsymbol{\alpha}$ e os expressam a partir dos resíduos de Pearson. No entanto, $\boldsymbol{\alpha}$ é estimado de maneira diferente de acordo com a

estrutura de correlação escolhida. De acordo com Hardin e Hilbe (2013), estimando $\boldsymbol{\alpha}$ a partir da estrutura de correlação apropriada para os dados e assim, consequentemente, a matriz de correlação de trabalho, ganha-se eficiência na estimativa dos parâmetros de regressão $\boldsymbol{\beta}$. A seguir serão apresentadas os estimadores de $\boldsymbol{\alpha}$ considerando as quatro diferentes estruturas de correlação discutidas na Seção 3.2, a independente, a permutável, a auto-regressiva de primeira ordem e a não estruturada.

1. **Independente**

   Nesse caso $\boldsymbol{\alpha}$ não precisa ser estimado, pois, como visto anteriormente, admite-se que ele assume o valor 0.

2. **Permutável**

$$\hat{\boldsymbol{\alpha}} = \frac{\hat{\phi}}{n} \sum_{i=1}^{n} \frac{1}{n_i(n_i-1)} \sum_{j=1}^{n_i} \sum_{\substack{j'=1 \\ j' \neq j}}^{n_i} r_{ij} r_{ij'}. \tag{3.20}$$

3. **Auto-regressiva de primeira ordem**

$$\hat{\boldsymbol{\alpha}} = \frac{\hat{\phi}}{n} \sum_{i=1}^{n} \frac{1}{(n_i-1)} \sum_{j=1}^{(n_i-1)} r_{ij} r_{i(j+1)}. \tag{3.21}$$

4. **Não estruturada**

$$\hat{\alpha}_{jj'} = \frac{\hat{\phi}}{n} \sum_{i=1}^{n} r_{ij} r_{ij'}. \tag{3.22}$$

3.3.5 Processo iterativo para estimação dos parâmetros

O algoritmo para o processo iterativo da estimação dos parâmetros pode ser dado através de

1. Calcular a estimativa inicial de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{(0)}$, como nos MLGs, assumindo independência entre as observações de um mesmo indivíduo.

2. A partir de $\hat{\boldsymbol{\beta}}^{(0)}$, calcular a estimativa do resíduo de Pearson, $\hat{r}_{ij}^{(0)}$.

3. A partir de $\hat{r}_{ij}^{(0)}$, calcular a estimativa do parâmetro de dispersão, $\hat{\phi}^{(0)}$, e dos parâmetros de correlação, $\hat{\boldsymbol{\alpha}}^{(0)}$.

4. Obter $\hat{\boldsymbol{\beta}}^{(1)}$ a partir da Expressão (3.15).

5. Repetir os passos 2. a 4. até a convergência.

O critério de parada utilizado para verificar a convergência pode ser o mesmo apresentado no Capítulo 2, mostrado na Equação (2.23).

3.4 Testes de hipóteses

Assim como visto na Seção 2.4, considere $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_a \\ \boldsymbol{\beta}_b \end{bmatrix}$ e o interesse é testar $H_0 : \boldsymbol{\beta}_b = \boldsymbol{\beta}_b^{(0)} \times H_1 : \boldsymbol{\beta}_b \neq \boldsymbol{\beta}_b^{(0)}$, em que $\boldsymbol{\beta}_b^{(0)}$ é um vetor de constantes conhecidas de dimensão K_b. Esse caso engloba tanto o teste para um único parâmetro como para vários, visto que $\boldsymbol{\beta}_b^{(0)}$ pode ser um vetor com apenas um escalar.

3.4.1 Teste de Wald generalizado

De acordo com Venezuela (2003), a adaptação do teste de Wald para equações de estimação generalizadas utiliza o estimador robusto da variância de $\hat{\boldsymbol{\beta}}$ em sua estatística teste, como podemos ver a seguir

$$Q_W^* = (\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^{(0)})^T \{[\widehat{\boldsymbol{V}}_{\boldsymbol{R}}]_{(b)}^{-1}\}^{-1} (\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^{(0)}), \tag{3.23}$$

em que $\hat{\boldsymbol{\beta}}_b$ é o estimador do parâmetro via EEGs sob H_0 e $[\widehat{\boldsymbol{V}}_{\boldsymbol{R}}]_{(b)}^{-1}$ é a submatriz da inversa de $\widehat{\boldsymbol{V}}_{\boldsymbol{R}}$ para $\hat{\boldsymbol{\beta}}_b$.

Sob H_0, Q_W^* tem distribuição assintótica qui-quadrado com K_b graus de liberdade. Então, rejeita-se H_0 se o valor observado da estatística for superior a $\chi^2_{K_b,1-\alpha}$, sendo α o nível de significância e $\chi^2_{K_b,1-\alpha}$ o quantil (1 - α) da $\chi^2_{K_b}$.

No entanto, Guo et al. (2005) destacam que quando o número de indivíduos é pequeno o teste não costuma funcionar bem. Diante disso, Rotnitzky e Jewell (1990) sugerem usar nesses casos uma estatística de Wald modificada que considera o estimador baseado no modelo para a variância de $\hat{\boldsymbol{\beta}}$, definido em (3.16), ao invés do robusto, com o intuito de que o tamanho do teste fique próximo a α.

3.5 Seleção de variáveis

O critério de informação de Akaike (AIC), como visto anteriormente na Seção 2.5, é um método amplamente utilizado para a seleção de modelos quando estamos tratando de MLGs. No entanto, ele não é aplicável as EEGs. Nesse contexto, Pan (2001) propôs o critério de quase-informação (QIC), que é uma modificação apropriada do AIC para as EEGs. Dito isto, o QIC pode ser expresso por

$$QIC = -2Q(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) + 2tr(\widehat{\boldsymbol{V}}_{\boldsymbol{M_I}}^{-\mathbf{1}}\widehat{\boldsymbol{V}}_{\boldsymbol{R}}), \tag{3.24}$$

em que $Q(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ é a função de quase-verossimilhança sob a hipótese de independência, sabendo que $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$ e $g^{-1}(.)$ é o inverso da função de ligação, onde $\boldsymbol{\beta}$ foi estimado considerando uma estrutura específica $\boldsymbol{R}(\boldsymbol{\alpha})$ de correlação. $\widehat{\boldsymbol{V}}_{\boldsymbol{M_I}}$ e $\widehat{\boldsymbol{V}}_{\boldsymbol{R}}$ são, respectivamente, a variância estimada baseada no modelo de $\hat{\boldsymbol{\beta}}$ sob a estrutura de independência (I) para a correlação e a variância robusta estimada de $\hat{\boldsymbol{\beta}}$ a partir de uma estrutura específica $\boldsymbol{R}(\boldsymbol{\alpha})$ para a correlação. Por fim, o termo de penalização é dado por $2tr(\widehat{\boldsymbol{V}}_{\boldsymbol{M_I}}^{-\mathbf{1}}\widehat{\boldsymbol{V}}_{\boldsymbol{R}})$.

Pan (2001) ainda sugeriu um método simplificado do QIC, denominado de QIC_u, que pode ser utilizado quando $tr(\widehat{\boldsymbol{V}}_{\boldsymbol{M_I}}^{-\mathbf{1}}\widehat{\boldsymbol{V}}_{\boldsymbol{R}}) \approx tr(\boldsymbol{I}) = p$ e é dado por

$$QIC_u = -2Q(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) + 2p, \tag{3.25}$$

em que p é o número de parâmetros do modelo e o termo de penalização é dado por $2p$.

Para selecionar o melhor modelo baseado nesses critérios apresentados, escolhe-se o que apresentar o menor valor do QIC ou QIC_u.

3.6 Seleção da estrutura de correlação

Fitzmaurice (1995) enfatiza que se a especificação da estrutura de correlação da matriz de trabalho, conhecida por $\boldsymbol{R}(\boldsymbol{\alpha})$, for incorreta pode afetar a eficiência das estimativas dos parâmetros de regressão. Sendo assim, é notória a importância da escolha da estrutura de correlação adequada. Ballinger (2004) sugere a escolha da correlação através do conhecimento da natureza dos dados. Nesse contexto, diz que se os dados forem longitudinais, é muito utilizada a correlação auto-regressiva, por considerar as correlações como uma função exponencial ao longo do tempo. Se for considerado que os dados estão agrupados em um assunto específico, é indicada a utilização da estrutura de correlação

permutável. Se for considerado que o número de observações de cada indivíduo é pequeno e os dados são balanceados, isto é, o número de observações de cada indivíduo é igual, é recomendada a correlação não estruturada. Por fim, se considerar a ausência de correlação nos dados, utiliza-se a estrutura de correlação independente. Vale ressaltar que apesar de ser possível escolher a estrutura de correlação a partir da natureza dos dados isso não é tão simples na prática, visto que, em geral, é preciso ser especialista no assunto referente aos dados para ter o conhecimento necessário para a escolha adequada. Dessa maneira, geralmente utilizam-se critérios estatísticos para selecionar a melhor estrutura de correlação para a matriz de trabalho.

O QIC pode ser um método utilizado para selecionar a melhor estrutura de correlação para $\boldsymbol{R}(\boldsymbol{\alpha})$. No entanto, segundo Cui (2007), o QIC_u não pode ser usado nesse caso, devido a suposição de equivalência assintótica de $\widehat{\boldsymbol{V}}_{\boldsymbol{M_I}}$ e $\widehat{\boldsymbol{V}}_{\boldsymbol{R}}$.

Hin e Wang (2009) propuseram um novo método para a seleção da estrutura da matriz de correlação de trabalho, baseado no termo de penalização do QIC, o qual foi denominado de critério de informação de correlação (CIC) e pode ser representado por

$$CIC = tr(\widehat{\boldsymbol{V}}_{\boldsymbol{M_I}}^{-\mathbf{1}}\widehat{\boldsymbol{V}}_{\boldsymbol{R}}). \tag{3.26}$$

Assim como no QIC, é escolhida a estrutura de correlação que apresentar o menor valor do CIC.

3.7 Análise de diagnóstico

No caso das EEGs, a detecção de pontos alavancas, discrepantes e influentes é feita de maneira similar aos MLGs, mas considerando os parâmetros de correlação nos cálculos das medidas descritas na Seção 2.6 do Capítulo anterior. Além disso, no que se diz respeito a avaliação da distribuição escolhida para a variável resposta, assim como nos MLGs, usualmente se faz com gráficos de probabilidade normal para um resíduo com a inclusão do envelope simulado. Porém, a diferença é que nas EEGs considera-se a matriz de correlação para gerá-los.

Novamente, vale ressaltar que como o objetivo do trabalhado é preditivo, usualmente não é feita análise de diagnóstico, visto que o modelo escolhido será o que apresentar o melhor poder de predição. Sendo assim, o leitor interessado em mais detalhes sobre análise de diagnóstico em EEGs pode consultar em Venezuela et al. (2007).

Capítulo 4

Aplicação

Neste capítulo é apresentado, na Seção 4.1, o banco de dados inicial que será utilizado. Na Seção 4.2 abordaremos a análise descritiva das variáveis e faremos algumas modificações no banco de dados. A modelagem dos dados é feita na Seção 4.3.

4.1 Banco de dados

O banco de dados foi criado a partir dos *scouts* do Cartola FC, tendo sido coletados dados de 31 atacantes da série A do Campeonato Brasileiro de 2019 nas 19 rodadas do segundo turno. É importante esclarecer que não foram incluídos no banco de dados todos os atacantes existentes, mas pelo menos está presente um jogador de cada um dos 20 clubes que disputam o campeonato, com o intuito de que haja representatividade no banco criado. Sendo assim, os atacantes escolhidos foram apenas os que apresentaram um bom desempenho e jogaram no mínimo metade das partidas do segundo turno. Os jogadores considerados, bem como seus clubes, podem ser vistos no Apêndice A. Dito isto, as variáveis do banco de dados são:

- **id:** número de identificação do jogador;
- **rod:** rodada referente a pontuação do jogador;
- **pont:** pontuação do jogador na rodada;
- **pont_ult:** pontuação do jogador na última rodada que ele disputou;
- **local:** mando de campo da partida disputada pelo jogador na rodada (sendo C caso a partida seja disputada em casa e F caso seja fora de casa);

- **gol:** quantidade de gols marcados pelo jogador na última rodada que ele disputou;
- **ass:** quantidade de assistências para gol dada pelo jogador na última rodada que ele disputou;
- **part:** participação em gols (gol + ass) do jogador na última rodada que ele disputou;
- **fd:** quantidade de finalizações defendidas pelo goleiro adversário dada pelo jogador na última rodada que ele disputou;
- **ff:** quantidade de finalizações pra fora dada pelo jogador na última rodada que ele disputou;
- **ftotal:** quantidade total de finalizações (fd + ff) dada pelo jogador na última rodada que ele disputou;
- **fs:** quantidade de faltas sofridas pelo jogador na última rodada que ele disputou;
- **fc:** quantidade de faltas cometidas pelo jogador na última rodada que ele disputou;
- **i:** quantidade de impedimentos do jogador na última rodada que ele disputou;
- **pe:** quantidade de passes errados do jogador na última rodada que ele disputou;
- **rb:** quantidade de roubadas de bola feitas pelo jogador na última rodada que ele disputou; e
- **cart:** quantidade total de cartões (vermelho + amarelo) recebidos pelo jogador na última rodada que ele disputou.

Como podemos perceber, a maioria das variáveis são de contagem (discretas), se referindo a quantidade de vezes que o jogador fez uma determinada ação, com exceção das variáveis **pont** e **pont_ult**, que são contínuas, e da variável **local**, que é qualitativa (*dummie*) e assume dois níveis. Ainda, vale ressaltar que a maioria dos jogadores não jogaram todas as 19 rodadas possíveis, com isso o número total de observações é de 485.

Esses dados podem ser considerados longitudinais, pois as variáveis de um mesmo indivíduo, neste caso os jogadores (**id**), são coletadas repetidas vezes ao longo do tempo, através de rodadas (**rod**). Com isso, justifica-se o uso das Equações de Estimação Generalizadas, pois esperamos que haja correlação entre as observações de um mesmo indivíduo. Nesse contexto, consideraremos a pontuação do jogador (**pont**) como a variável resposta, tendo como foco tentar predizer os seus valores, e as demais variáveis como preditoras.

4.2 Análise descritiva

Todas as estatísticas descritivas e gráficos apresentados ao longo dessa seção foram feitos no *software* R, bem como os testes de hipóteses, os quais consideraremos um nível de significância de 5% para efetuar as análises. No Apêndice B pode ser visto a programação.

4.2.1 Variável resposta

Ao longo dessa subseção faremos um estudo da variável resposta.

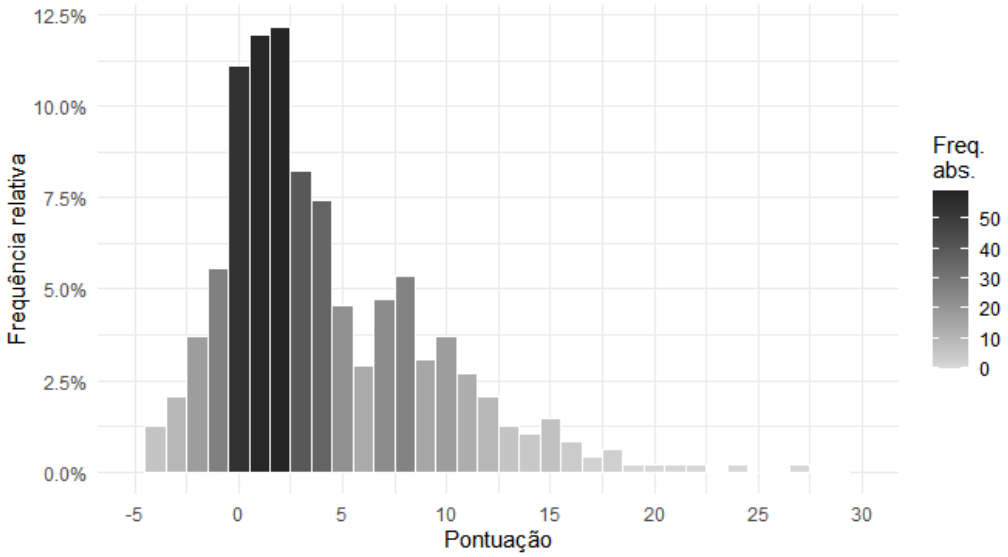

Figura 4.1: Histograma da pontuação do jogador na rodada.

A Figura 4.1 mostra o histograma da variável resposta. Dessa forma, percebe-se que os valores mais frequentes da variável se encontram no intervalo de -1 a 3. Além disso, os dados não aparentam se comportar de maneira simétrica em torno da média, mas não descartaremos a possibilidade de que a distribuição da variável resposta seja normal na modelagem dos dados, visto que devemos considerar a distribuição de $Y|X$ e não de Y simplesmente. Ainda, percebe-se uma assimetria à direita (positiva) dos dados, o que nos sugere que a pontuação dos jogadores possa seguir uma distribuição gama ou gaussiana inversa, que são positivas, contínuas e assimétricas. Porém, é visível que a variável apresenta valores negativos, sendo assim, para trabalharmos com essas distribuições teremos que fazer uma transformação nela para que seu espaço amostral contenha apenas valores positivos. Diante desse problema analisaremos a seguir as medidas resumo da variável.

Vale destacar ainda que assumir que a distribuição de $Y|X$ é normal apresenta a vantagem de podermos trabalhar com a variável resposta da forma que ela foi medida, o

que não é possível se assumirmos que a distribuição de $Y|X$ seja alguma das distribuições assimétricas mais conhecidas.

Tabela 4.1: Medidas resumo da pontuação do jogador na rodada.

Mínimo	**1° quartil**	**Mediana**	**Média**	**Variância**	**3° quartil**	**Máximo**
-4.9	0.6	2.7	4.154	25.452	7.4	27.5

A Tabela 4.1 apresenta as medidas resumo da variável resposta. Sendo assim, vemos que o valor mínimo que a variável apresenta é de -4.9. Então, para que as distribuições gama e gaussiana inversa possam ser consideradas para a variável resposta, acrescentaremos 5 em todos os seus valores, dando origem a uma nova variável, a qual chamaremos de pontuação modificada dos jogadores (**pont_mod**). É importante ressaltar que, como o objetivo do trabalho é preditivo, isso não traz consequências indesejáveis. Além disso, pode-se dizer que 50% dos valores da pontuação do jogador na rodada estão concentrados no intervalo de 0.6 a 7.4. Por fim, vemos que a diferença do máximo e o 3° quartil é bem maior que a do 1° quartil e o mínimo, o que é mais um indício da assimetria nos valores da variável.

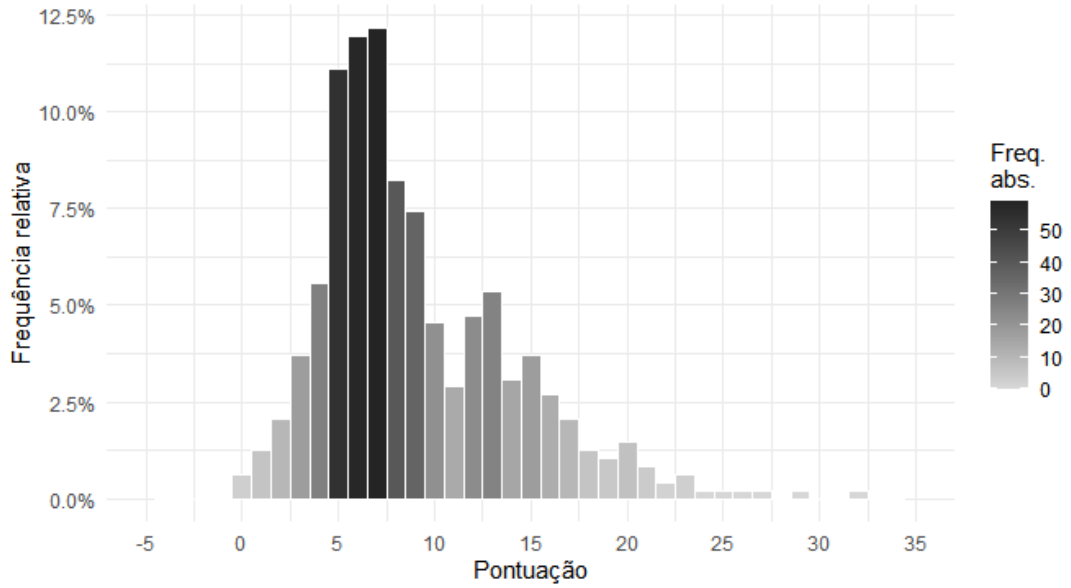

Figura 4.2: Histograma da pontuação modificada do jogador na rodada.

A Figura 4.2 apresenta o histograma da variável resposta após somar o valor 5 em todas as suas observações. Dessa maneira, vemos que agora todos os valores da variável são positivos.

Como a intenção é aplicar EEG, vamos verificar se realmente é razoável considerar a presença de correlação entre a pontuação modificada do jogador na rodada e a pontuação

modificada do jogador na última rodada que ele disputou, que é a variável **pont_ult** acrescida de 5 em todos os seus valores, a qual denominaremos de **pont_ult_mod**. Isso será feito porque, como vimos ao longo do Capítulo 3, as EEGs consideram a existência de correlação entre as observações de um mesmo indivíduo. Tendo em vista que a variável **pont_ult_mod** é nada mais do que a variável **pont_mod** na observação anterior, essa é uma maneira de verificar se a metodologia realmente é adequada para o nosso caso. Com isso, analisaremos a seguir 3 diferentes coeficientes de correlação entre essas variáveis, o de Pearson, o de Spearman e o de Kendall e seus respectivos testes de correlação que avaliam as hipóteses H_0: as variáveis não são correlacionadas x H_1: as variáveis são correlacionadas.

Tabela 4.2: Coeficientes de correlação referente às variáveis pontuação modificada do jogador na rodada e pontuação modificada do jogador na última rodada que ele disputou.

Método	**Coeficiente de correlação**
Pearson	0.0884
Spearman	0.0929
Kendall	0.0626

Tabela 4.3: Testes de correlação referente às variáveis pontuação modificada do jogador na rodada e pontuação modificada do jogador na última rodada que ele disputou.

Teste de correlação	**P-valor**
Pearson	0.0597
Spearman	0.0477
Kendall	0.0477

A Tabela 4.2 expõe os coeficientes de correlação entre a pontuação modificada do jogador na rodada e pontuação modificada do jogador na última rodada que ele disputou, enquanto a Tabela 4.3 apresenta os seus respectivos testes de correlação. Dessa forma, como podemos perceber através da Tabela 4.2 o coeficiente de correlação é baixo nos 3 casos. Agora, de acordo com a Tabela 4.3, pode-se concluir que, a partir dos testes de correlação de Spearman e de Kendall, há evidências de que as variáveis possuem correlação, pois rejeitamos H_0, visto que os seus p-valores são menores que o nível de significância, mas vale ressaltar que eles estão no limítrofe para não rejeitar. Por outro lado, o teste de correlação de Pearson não rejeita H_0, e com ele temos a conclusão de que não há evidência de que haja correlação entre as variáveis. Esses resultados apontados podem ser explicados devido ao fato de que, em geral, um jogo em casa é seguido de um jogo fora de casa e a tendência é que os jogadores pontuem mais em jogos em casa

e menos em jogos fora de casa, sendo assim, há uma grande oscilação na pontuação dos jogadores, fazendo com que, consequentemente, apresente uma alta variabilidade, como vimos na Tabela 4.1. No entanto, ainda faz todo sentido que haja essa correlação.

Diante do problema exposto, agruparemos todas as variáveis de 3 em 3 rodadas, o qual será feito da seguinte forma: somaremos até 3 valores de todas as variáveis de um mesmo indivíduo em rodadas subsequentes e faremos a sua média, sendo que para isso o jogador tenha que ter jogado pelo menos 1 partida no período de 3 rodadas consecutivas.

É importante se atentar ao fato de que agora as variáveis preditoras, com exceção da **local** que seguirá alinhada à resposta, começarão na rodada 20 e irão até a 34, cujos agrupamentos começarão então da rodada 20 até a 22, depois da 23 até a 25, até que chegue na último que vai da rodada 32 até a 34, enquanto a variável resposta vai da rodada 23 até a 37, em que o primeiro agrupamento será referente à rodada 23 até a 25 e o último à rodada 35 até a 37. Com isso, a rodada 38 não será utilizada no banco de dados novo, que passará a ter 154 observações e pode ser visto no Apêndice A.

Todas as análises descritivas feitas posteriormente considerarão as variáveis do banco de dados novo, o qual, por conveniência, denominaremos o nome de suas variáveis de maneira equivalente aos apresentados na Seção 4.1.

Nesse contexto, a nova variável resposta, a qual chamaremos de pontuação agrupada do jogador, será criada fazendo a média das pontuações dos jogadores de 3 em 3 rodadas subsequentes, a partir da variável **pont_mod**.

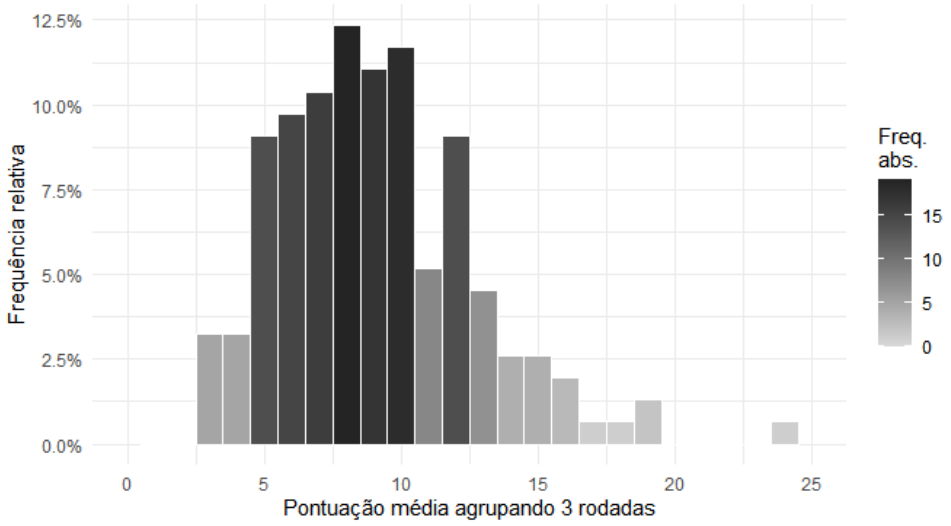

Figura 4.3: Histograma da pontuação agrupada do jogador.

A Figura 4.3 apresenta o histograma da variável resposta considerando o agrupamento. Dessa forma, vemos que a maioria das observações estão concentrados no intervalo

de 5 a 10 pontos. Além disso, aparentemente, os dados sugerem uma assimetria positiva, mas, novamente, não descartaremos a possibilidade de que a distribuição da variável resposta seja normal na hora de modelar os dados, visto que devemos considerar a distribuição de $Y|X$ e não de Y simplesmente. Apesar disso, mais uma vez, há indícios de que distribuições como a gama e a gaussiana inversa se adequariam melhor a variável resposta do que a distribuição normal.

Tabela 4.4: Medidas resumo da pontuação agrupada do jogador.

Mínimo	**1° quartil**	**Mediana**	**Média**	**Variância**	**3° quartil**	**Máximo**
2.94	6.515	8.7	9.144	12.84	11.457	23.7

A Tabela 4.4 apresenta as medidas resumo da variável resposta considerando o agrupamento. Dessa maneira, conseguimos extrair que 50% dos valores da pontuação agrupada do jogador estão concentrados no intervalo de 6.515 e 11.457. Além disso, é importante dizer que a variância da pontuação do jogador decaiu de 25.45 para 12.84 ao agruparmos os dados. Por fim, vemos que a diferença do máximo e o 3° quartil é bem maior que a do 1° quartil e o mínimo, o que é, novamente, mais um indício da assimetria nos valores da variável.

Agora, novamente analisaremos 3 diferentes coeficientes de correlação e seus respectivos testes, referentes a pontuação do jogador e à pontuação na última rodada que ele disputou para justificarmos a utilização das EEGs. Porém, dessa vez consideramos essas variáveis do banco de dados novo, que considera o agrupamento dos dados de 3 em 3 rodadas, como foi dito anteriormente. Sendo assim, a nova variável que representa a pontuação na última rodada que ele disputou no banco de dados novo, daremos o nome de pontuação agrupada do jogador nas últimas 3 rodadas. Essa nova variável será criada fazendo a média da pontuação do jogador nas 3 rodadas anteriores a de início da variável resposta. Caso o jogador não tenha jogado nenhuma das 3 partidas referentes as rodadas será considerada a pontuação agrupada das últimas 3 rodadas em que o jogador jogou em pelo menos uma partida.

Tabela 4.5: Coeficientes de correlação referente as variáveis pontuação agrupada do jogador e pontuação agrupada do jogador nas últimas 3 rodadas.

Método	**Coeficiente de correlação**
Pearson	0.3300
Spearman	0.2764
Kendall	0.1863

Tabela 4.6: Testes de correlação referente as variáveis pontuação agrupada do jogador e pontuação agrupada do jogador nas últimas 3 rodadas.

Teste de correlação	**P-valor**
Pearson	< 0.0001
Spearman	0.0005
Kendall	0.0006

A Tabela 4.5 expõe os coeficientes de correlação entre a pontuação agrupada do jogador e pontuação agrupada do jogador nas últimas 3 rodadas, enquanto a Tabela 4.6 apresenta os seus respectivos testes de correlação. Dessa forma, analisando a Tabela 4.5 vemos que os coeficientes de correlação aumentaram consideravelmente em relação aos dados sem considerar o agrupamento, apesar de ainda não serem altos, sendo o maior o de Pearson com 0.33 de correlação entre as variáveis. Além disso, em todos os testes de correlação apresentados na Tabela 4.6, o valor-p é desprezível e menor que o nível de significância, ou seja, há evidências de que existe a correlação entre as variáveis. Diante do exposto, apesar de não serem altos os coeficientes de correlação, é natural considerar a presença de correlação entre as observações. Dessa maneira, justifica-se o uso das EEGs.

Antes de iniciarmos a análise descritiva das variáveis preditoras, é importante dizer que o banco de dados foi dividido em duas partes. A primeira, tendo em vista a variável resposta, é composta das observações referentes aos grupos de rodadas 23 a 25, 26 a 28 e 29 a 31 e 32 a 34. No entanto, vale ressaltar que o agrupamento das rodadas 20 a 22 estão incluídos nessa divisão, mas apenas como variáveis preditoras das observações em que a variável resposta é do agrupamento das rodadas 23 a 25. A exceção se dá pela variável mando de campo, que é a única preditora em que agrupamento de referência é igual ao da variável resposta. Essa primeira parte da divisão dos dados (base de dados de treino), que contará com 123 observações, é a que será utilizada para as análises descritivas posteriores e no ajuste dos modelos. A segunda parte, é composta apenas por uma observação de cada jogador, e é constituída pelas observações em que a variável resposta está contida no agrupamento das rodadas 35 a 37 e, assim, consequentemente, as preditoras, com exceção da variável mando de campo, são do agrupamento das rodadas 32 a 34. A avaliação do poder preditivo dos modelos ajustados será feita a partir dessa segunda parte dos dados (base de dados de teste), que contará com 31 observações.

4.2.2 Histogramas

Ao longo dessa Subseção serão apresentados os histogramas de todas as variáveis preditoras do banco de dados novo e suas respectivas análises.

Vale ressaltar que, exceto a pontuação agrupada do jogador nas últimas 3 rodadas que ele disputou, quase todas as outras variáveis preditoras passam a assumir apenas determinados valores, como se fossem "fatores", por conta de originalmente serem variáveis quantitativas discretas. Isso acontece pois, quando é feito o agrupamento dos valores das variáveis em 3 rodadas, a princípio temos a quantidade acumulada de determinada estatística do jogador e, assim, posteriormente, dividindo as mesmas pela quantidade de jogos disputados pelo jogador no grupo de rodadas em questão, obtemos os seus valores médios. De maneira análoga, no caso do mando de campo, que inicialmente era uma variável qualitativa que assumia dois níveis, passa a ser uma variável quantitativa, em que seus valores são a porcentagem de jogos que o jogador disputou dentro de casa no agrupamento das rodadas. Por fim, em relação a pontuação agrupada do jogador nas últimas 3 rodadas que ele disputou, por ser originalmente uma quantitativa contínua, não ocorre essa restrição dos valores de suas observações.

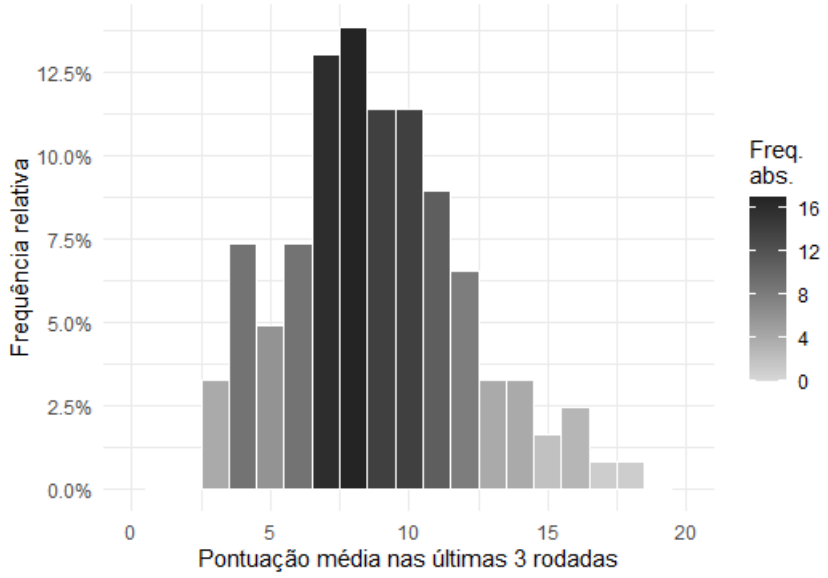

Figura 4.4: Histograma da pontuação agrupada do jogador nas últimas 3 rodadas.

A Figura 4.4 apresenta o histograma da pontuação agrupada do jogador nas últimas 3 rodadas que ele disputou. Dessa maneira, observamos que os valores mais frequentes da variável estão entre 7 e 10 pontos, sendo que estes, juntos, representam quase metade da frequência da variável.

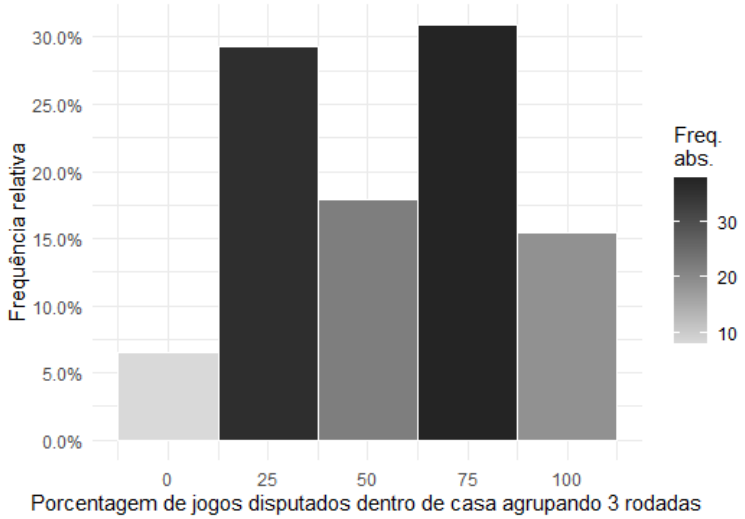

Figura 4.5: Histograma do mando de campo agrupado de partidas disputadas pelo jogador.

A Figura 4.5 traz o histograma do mando de campo agrupado de partidas disputadas pelo jogador. Vemos que em cerca de 15% dos casos o jogador disputou 100% dos jogos dentro de casa, considerando o agrupamento de 3 rodadas. Há observações com 50% dos jogos em casa pois, conforme já mencionado, para todas as variáveis consideram-se apenas os jogos que o jogador participou, ou seja, neste caso, por exemplo, o atacante jogou apenas 2 partidas no período de 3 rodadas, sendo uma dentro e outra fora de casa.

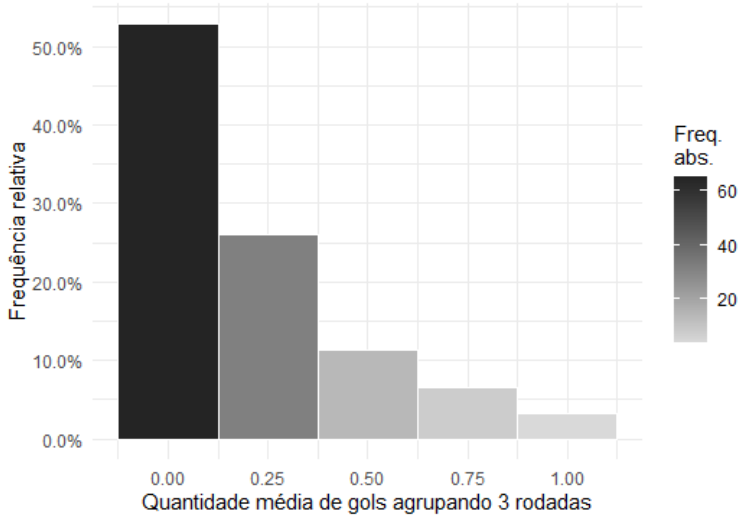

Figura 4.6: Histograma da quantidade agrupada de gols marcados pelo jogador.

A Figura 4.6 apresenta o histograma da quantidade agrupada de gols marcados pelo jogador. Da mesma, vemos que a maioria das observações apresentam uma quantidade média de 0 gols considerando o agrupamento de rodadas. Além disso, observamos que, quanto maior a média de gols, menor é a frequência relativa.

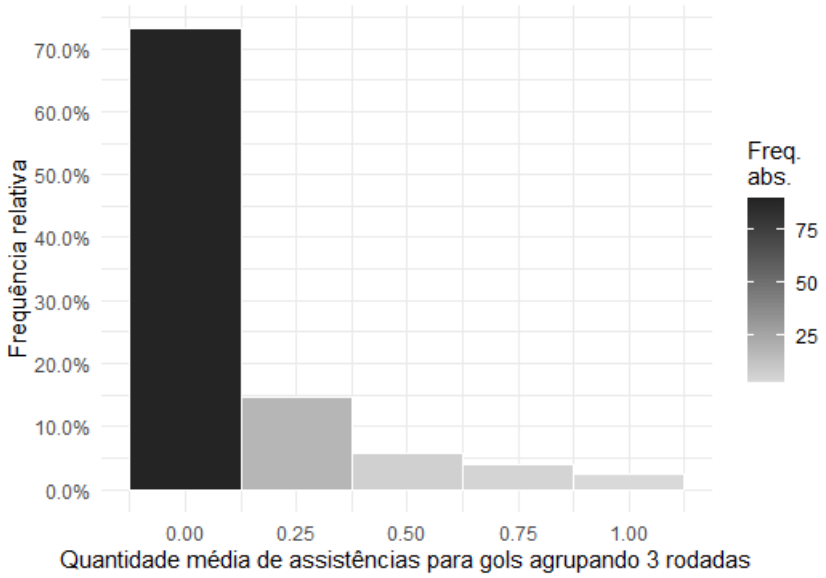

Figura 4.7: Histograma da quantidade agrupada de assistências para gol dadas pelo jogador.

A Figura 4.7 traz o histograma da quantidade agrupada de assistências para gol dadas pelo jogador. Dessa maneira, analisamos que mais de 70% das observações assumem o valor 0 para a quantidade média de assistências para gols agrupando 3 rodadas.

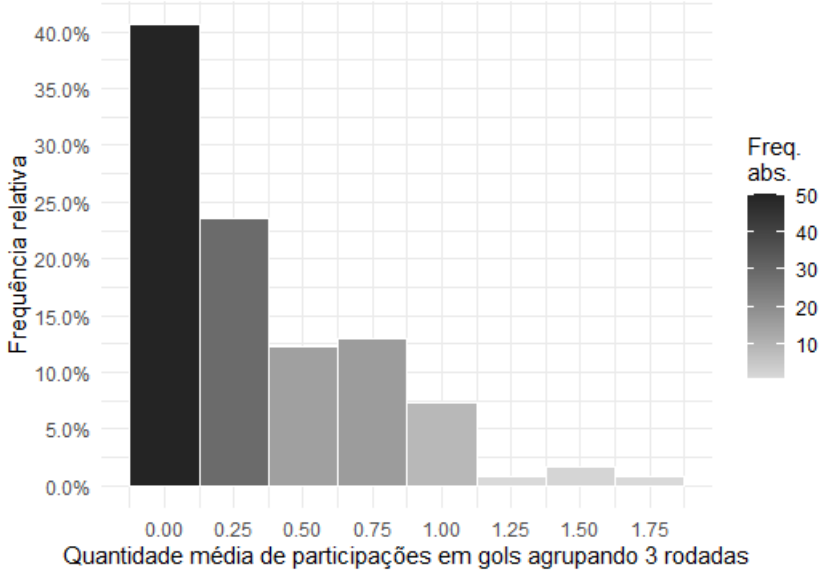

Figura 4.8: Histograma da quantidade agrupada de participações em gols do jogador.

A Figura 4.8 apresenta o histograma da quantidade agrupada de participações em gols do jogador. A partir dela concluímos que em cerca de 60% das observações o jogador participou, em média, de 0.33 ou mais gols no período de 3 rodadas.

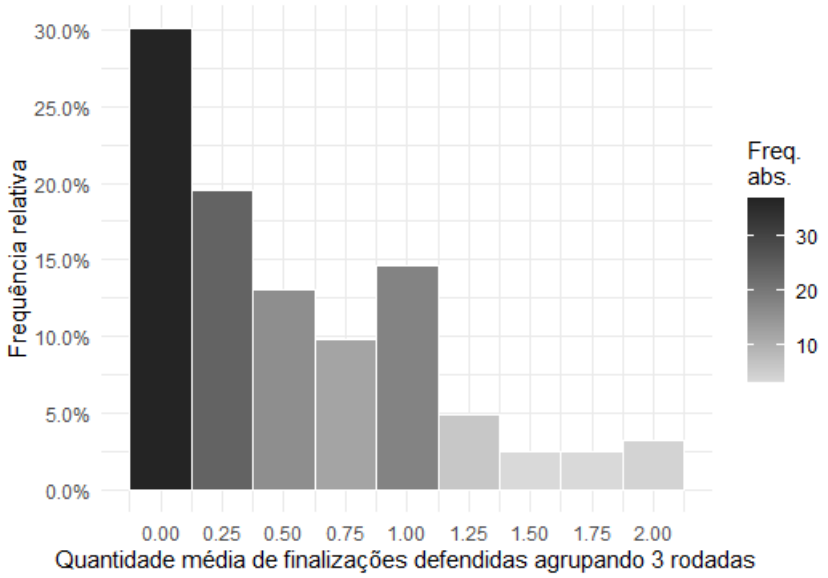

Figura 4.9: Histograma da quantidade agrupada de finalizações defendidas pelo goleiro adversário dadas pelo jogador.

A Figura 4.9 traz o histograma da quantidade agrupada de finalizações defendidas pelo goleiro adversário dadas pelo jogador. Dela, temos que em cerca de 70% das observações, considerando o período de 3 rodadas, o jogador fez, em média, 0.33 ou mais finalizações que foram defendida pelo goleiro adversário.

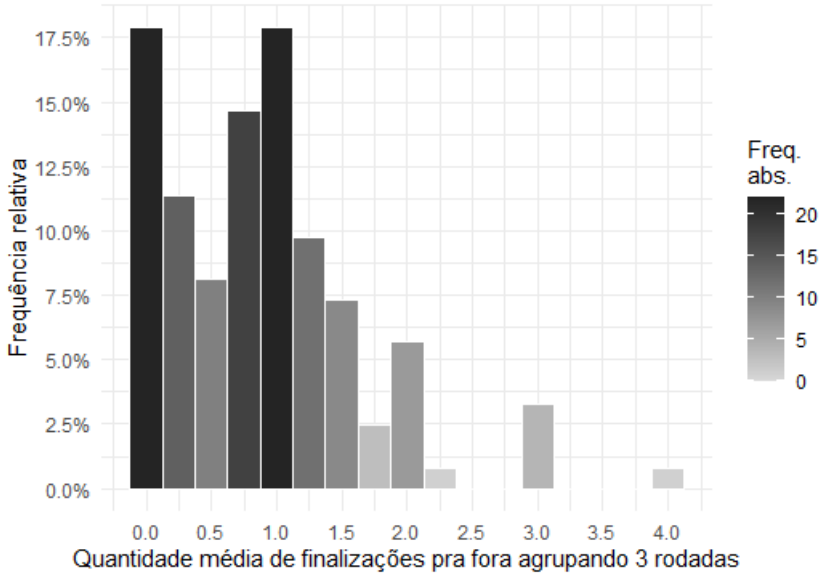

Figura 4.10: Histograma da quantidade agrupada de finalizações pra fora dadas pelo jogador.

A Figura 4.10 apresenta o histograma da quantidade agrupada de finalizações pra fora dadas pelo jogador. A partir da mesma, podemos dizer que em apenas cerca de 18% das observações o jogador não finalizou pra fora nenhuma vez no período de 3 rodadas.

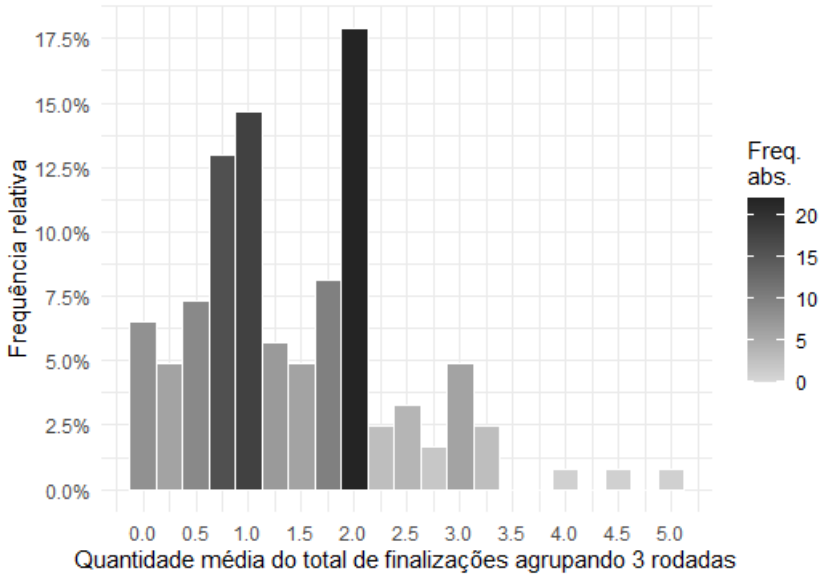

Figura 4.11: Histograma da quantidade agrupada do total de finalizações dadas pelo jogador.

A Figura 4.11 traz o histograma da quantidade agrupada do total de finalizações dadas pelo jogador. Dessa maneira, observamos que em cerca de 70% das observações o total de finalizações dadas pelo jogador foi, em média, considerando o agrupamento de 3 rodadas, pelo menos um em cada partida disputada.

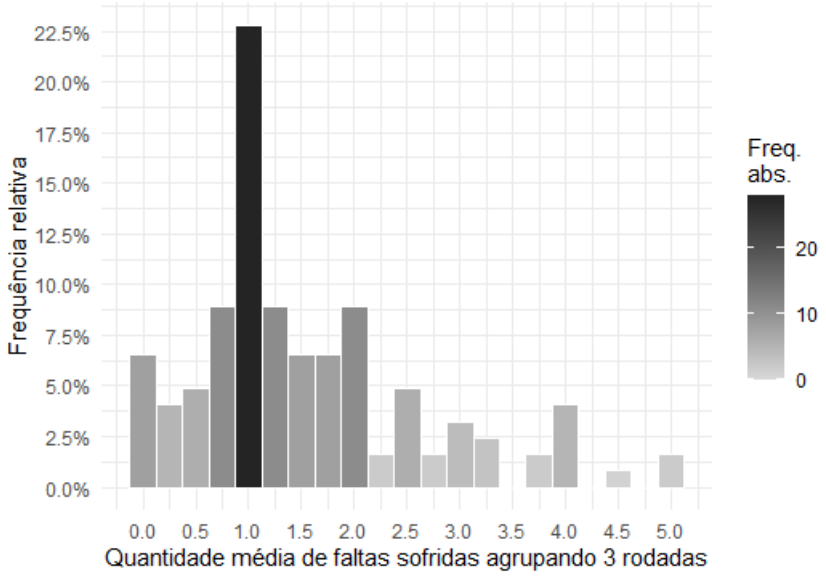

Figura 4.12: Histograma da quantidade agrupada de faltas sofridas pelo jogador.

A Figura 4.12 apresenta o histograma da quantidade agrupada de faltas sofridas pelo jogador. Dessa maneira, concluímos que em cerca de 23% das observações o jogador sofreu em média uma falta por jogo disputado no período de 3 rodadas.

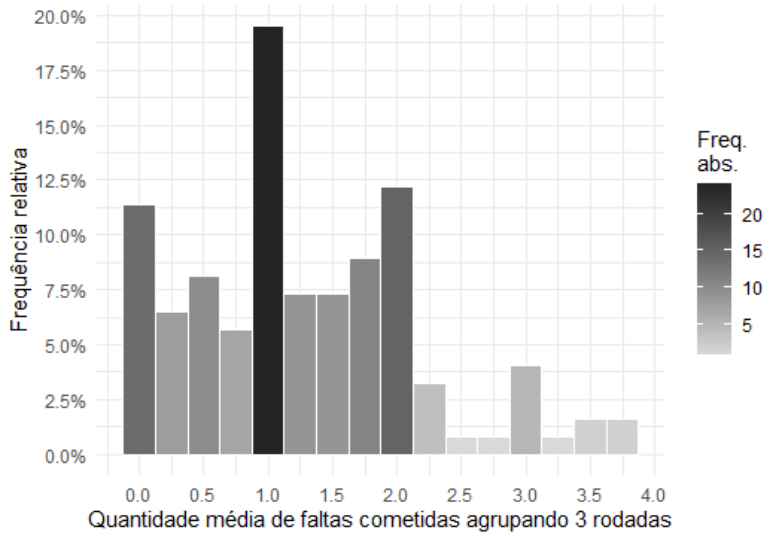

Figura 4.13: Histograma da quantidade agrupada de faltas cometidas pelo jogador.

A Figura 4.13 traz o histograma da quantidade agrupada de faltas cometidas pelo jogador. A partir dela, concluímos que por volta de 19% das observações o jogador cometeu em média uma falta no período de 3 rodadas em cada partida disputada, sendo o valor mais frequente dentre os assumidos pela variável. Por outro lado, para cerca de 11% das observações a média de faltas é zero.

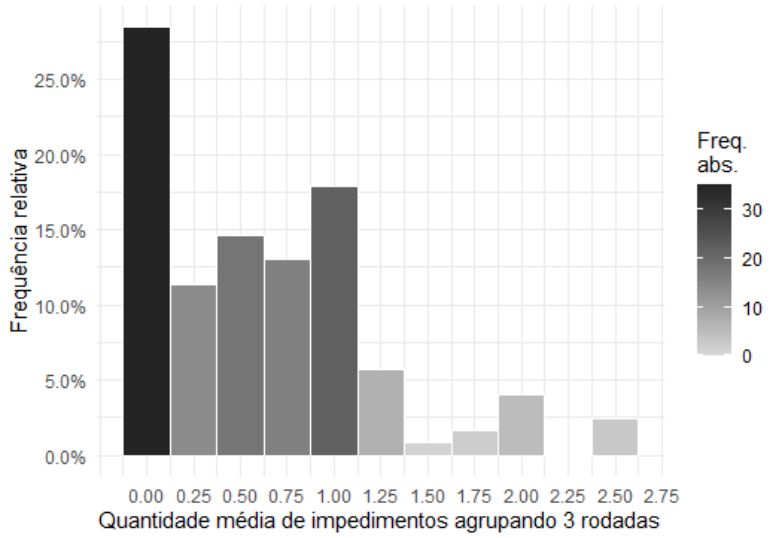

Figura 4.14: Histograma da quantidade agrupada de impedimentos do jogador.

A Figura 4.14 apresenta o histograma da quantidade agrupada de impedimentos do jogador. Dessa forma, vemos que por volta de 28% das observações a quantidade média de impedimentos é 0 agrupando 3 rodadas. Por outro lado, em cerca de 18% das observações o jogador apresenta uma média de 1 impedimento por partida jogada agrupando 3 rodadas.

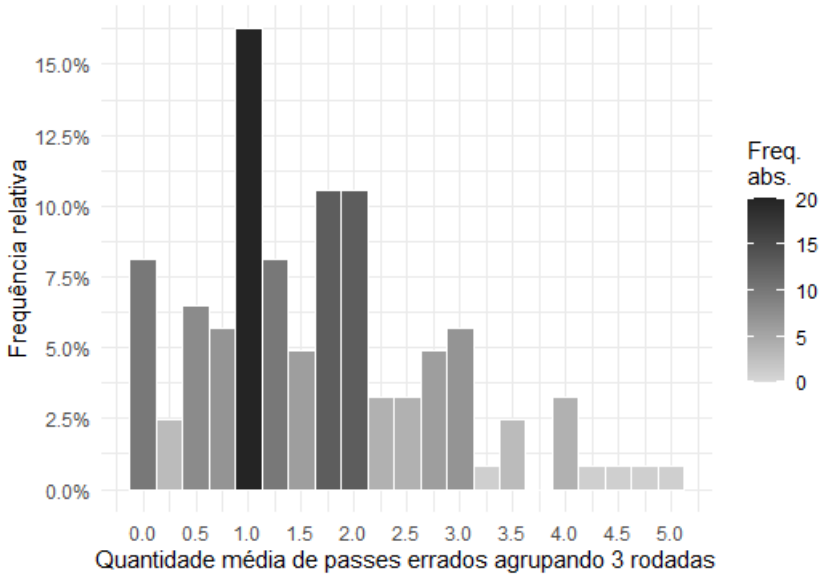

Figura 4.15: Histograma da quantidade agrupada de passes errados dados pelo jogador.

A Figura 4.15 traz o histograma da quantidade agrupada de passes errados dados pelo jogador. Da mesma forma, analisamos que em apenas cerca de 8% das observações o jogador não errou nenhum passe no período de 3 rodadas, enquanto em mais de 50% das observações o jogador errou entre 1 e 2 passes em média.

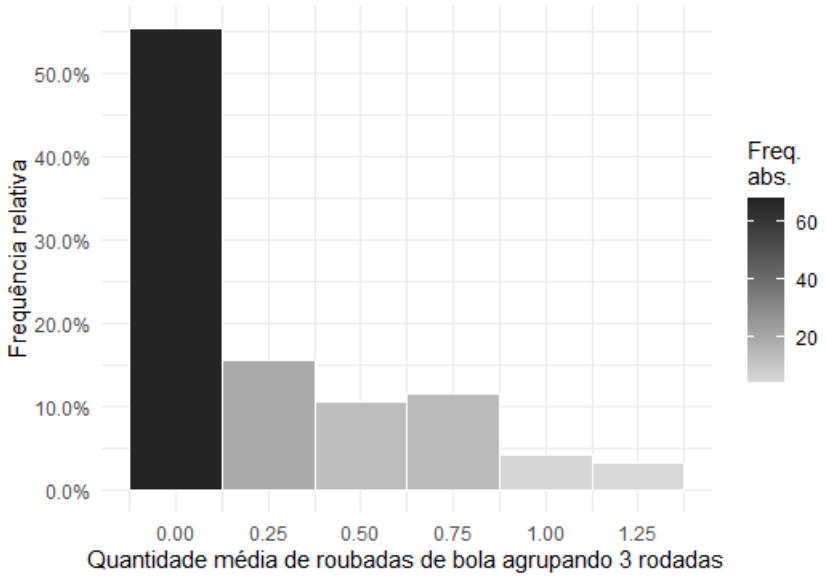

Figura 4.16: Histograma da quantidade agrupada de roubadas de bola feitas pelo jogador.

A Figura 4.16 apresenta o histograma da quantidade agrupada de roubadas de bola feitas pelo jogador. A partir dela, observamos que em mais da metade das observações o jogador não roubou nenhuma bola no período de 3 rodadas.

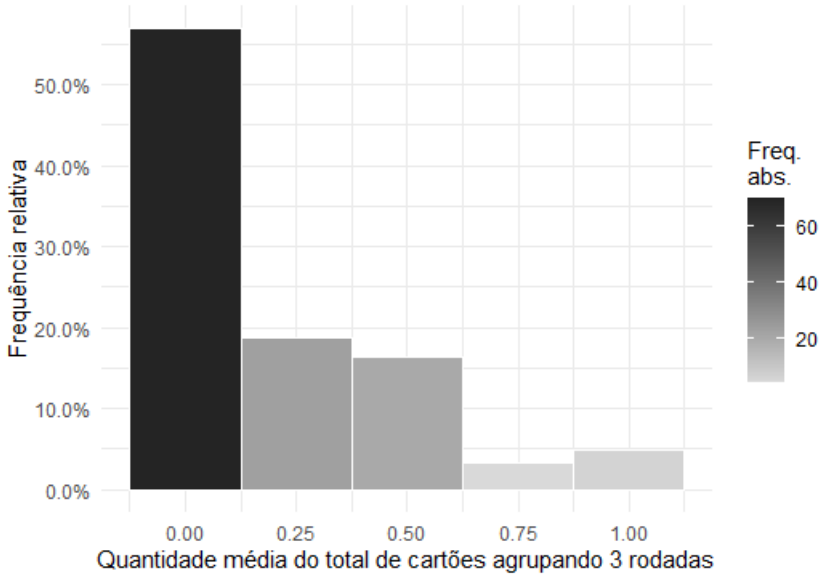

Figura 4.17: Histograma da quantidade agrupada do total de cartões recebidos pelo jogador.

A Figura 4.17 traz o histograma da quantidade agrupada do total de cartões recebidos pelo jogador. Dessa maneira, vemos que em mais da metade das observações o jogador não recebeu nenhum cartão, seja este amarelo ou vermelho, considerando o agrupamento de 3 rodadas.

4.2.3 Diagramas de dispersão

Nessa subseção apresentaremos todos os diagramas de dispersão das variáveis preditoras versus a variável resposta.

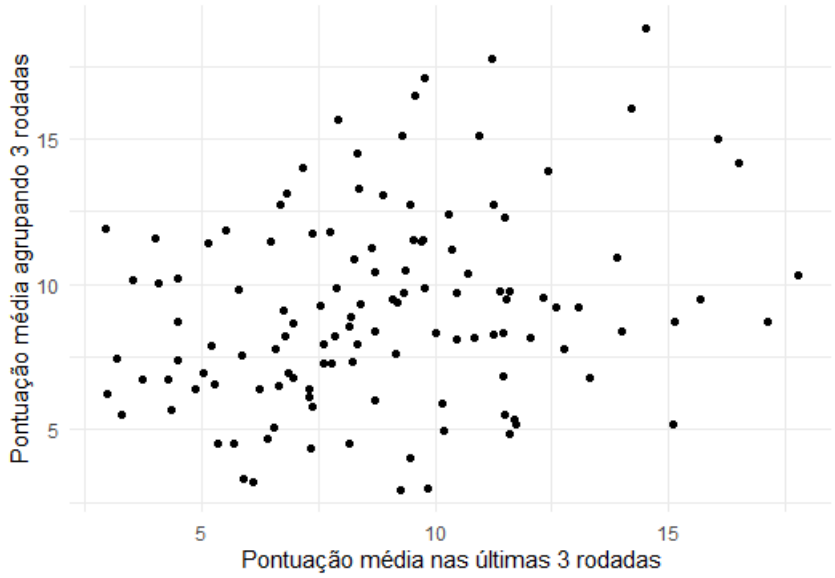

Figura 4.18: Diagrama de dispersão da pontuação agrupada do jogador nas últimas 3 rodadas versus a pontuação agrupada do jogador.

A Figura 4.18 apresenta o diagrama de dispersão da pontuação agrupada do jogador nas últimas 3 rodadas versus a pontuação agrupada do jogador. Dessa maneira, observamos que há indícios de existir uma leve correlação positiva entre as duas variáveis. Isso nos leva a crer que, quando for feita a modelagem dos dados, a variável preditora pontuação agrupada do jogador nas últimas 3 rodadas deve ser significativa para o estudo da variável resposta pontuação agrupada do jogador.

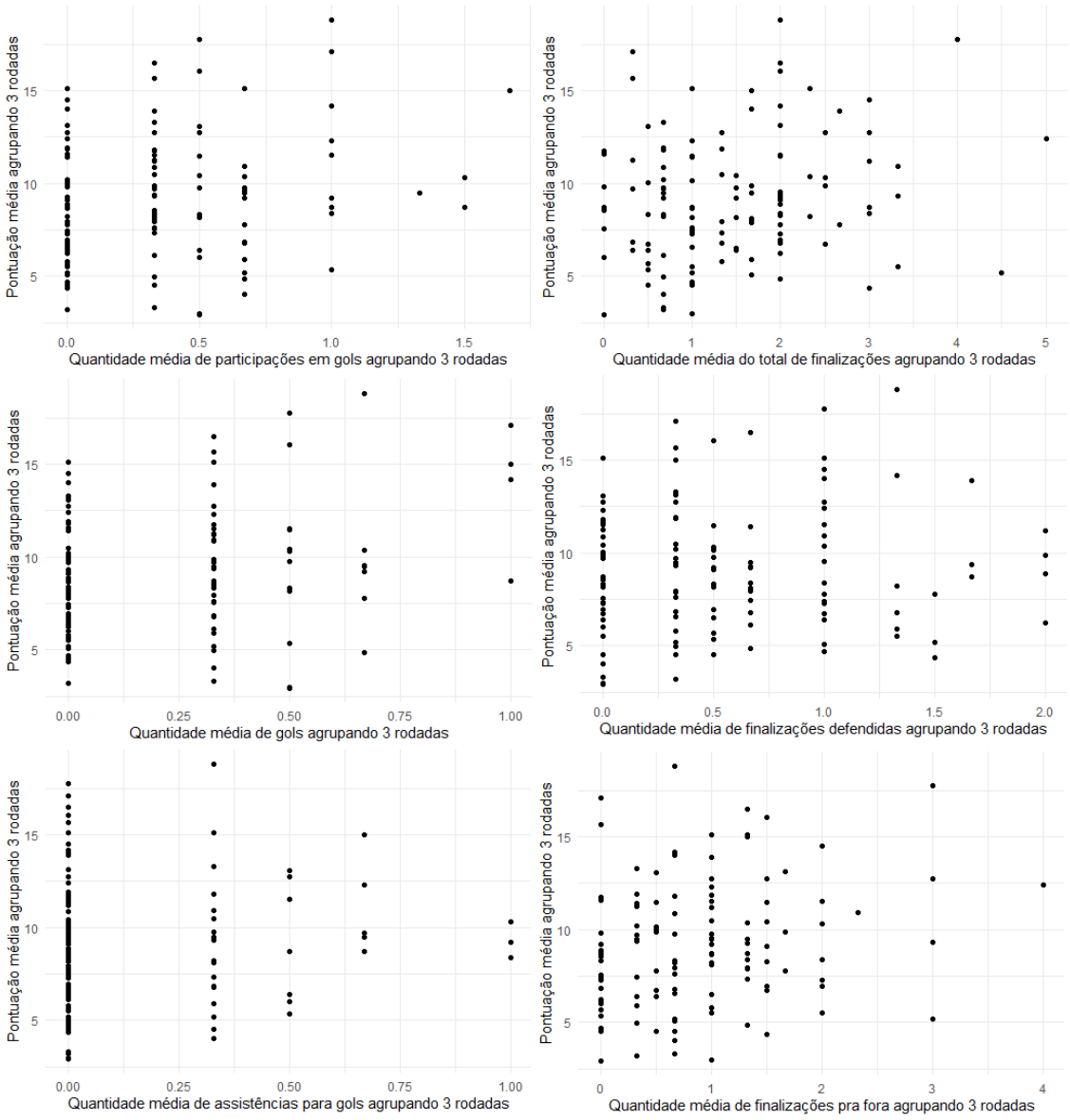

Figura 4.19: Diagramas de dispersão das variáveis preditoras relacionadas a finalizações e envolvimento em gols versus a pontuação agrupada do jogador.

A Figura 4.19 apresenta os diagramas de dispersão das variáveis preditoras relacionadas a finalizações e envolvimento em gols versus a pontuação agrupada do jogador. Dessa forma, primeiramente, podemos ver que as variáveis que foram criadas a partir de outras, ou seja, a quantidade agrupada de participações em gols do jogador e a quantidade agrupada do total de finalizações dadas pelo jogador, apresentam mais níveis do que suas respectivas derivadas. Além disso, vemos que, em cada um dos valores assumidos pelas variáveis preditoras, há uma grande variação nos valores da variável resposta. Dessa maneira, diante do exposto, para uma melhor avaliação da associação entre cada variável preditora e a resposta, faremos diagramas de dispersão considerando a média da variável resposta para cada valor que as variáveis preditoras assumem.

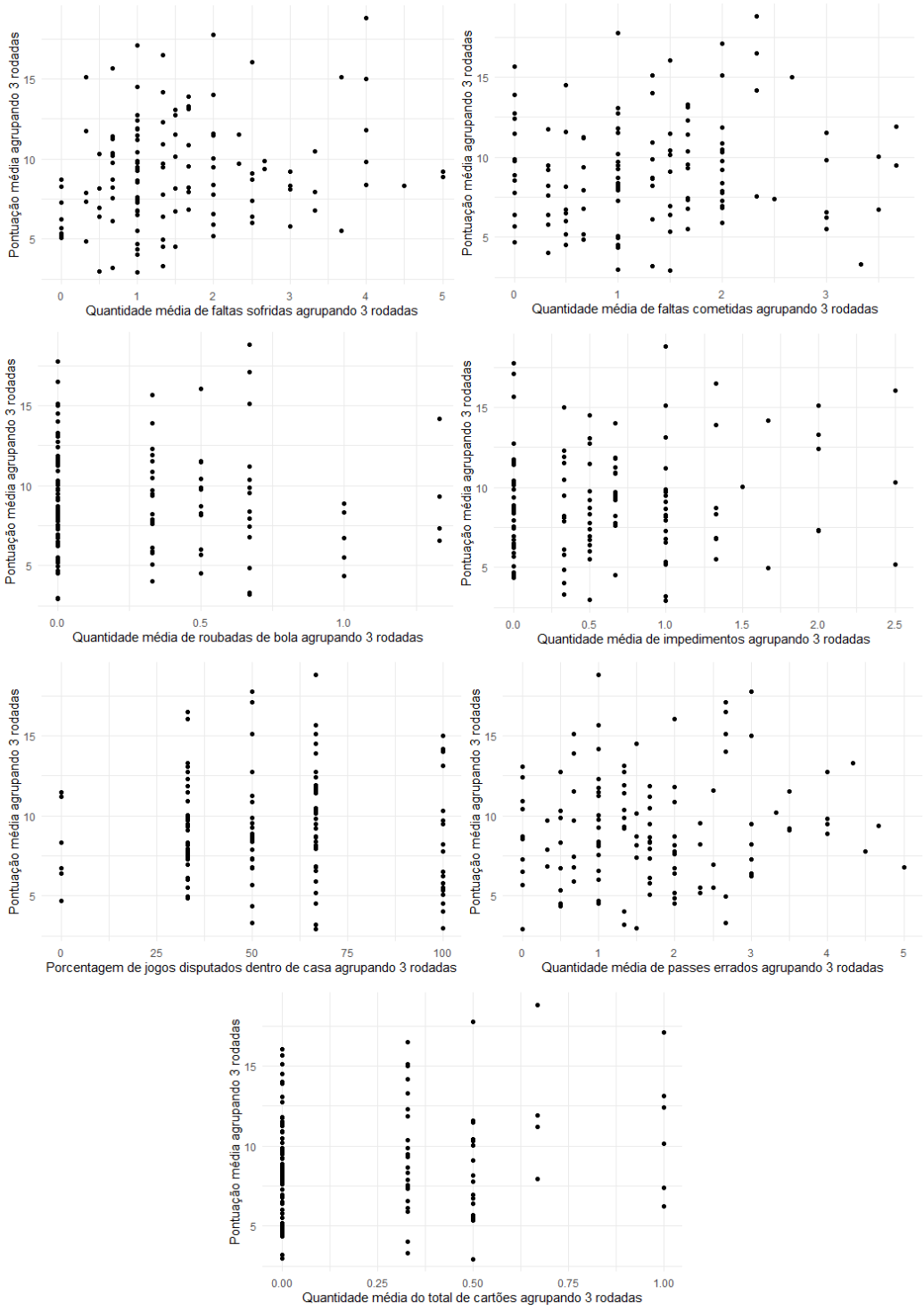

Figura 4.20: Diagramas de dispersão das variáveis preditoras restantes versus a pontuação agrupada do jogador.

A Figura 4.20 traz os diagramas de dispersão das variáveis preditoras restantes versus a pontuação agrupada do jogador. Podemos perceber que ocorre o mesmo problema da variação nos valores da variável resposta em cada valor assumido pelas variáveis preditoras. Sendo assim, também faremos para essas variáveis preditoras diagramas de dispersão considerando a média da variável resposta para cada valor que elas assumem.

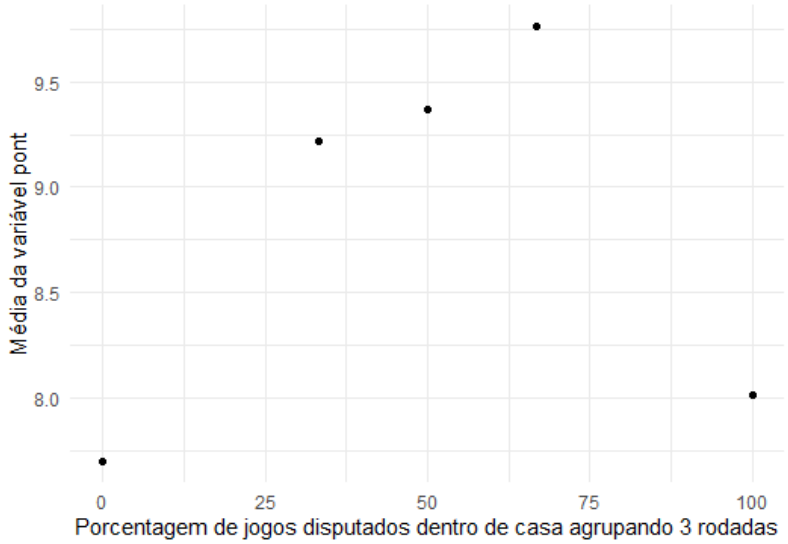

Figura 4.21: Diagrama de dispersão do mando de campo agrupado de partidas disputadas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.21 apresenta o diagrama de dispersão do mando de campo agrupado de partidas disputadas pelo jogador versus a média da pontuação agrupada do jogador. Observa-se que as variáveis não parecem ter associação linear.

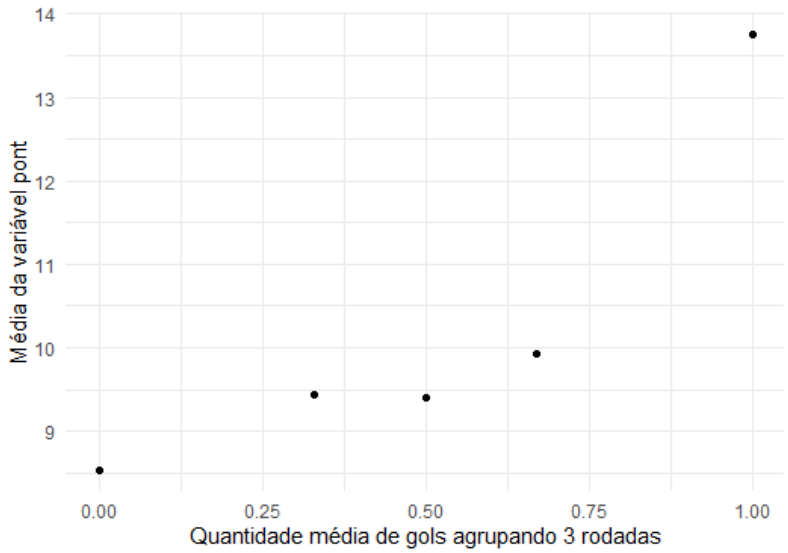

Figura 4.22: Diagrama de dispersão da quantidade agrupada de gols marcados pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.22 traz o diagrama de dispersão da quantidade agrupada de gols marcados pelo jogador versus a média da pontuação agrupada do jogador. Dessa maneira, analisamos que, em geral, parece que, conforme aumentamos o valor da variável preditora, maior é a média da variável resposta, o que é um indício de que pode haver correlação entre as duas variáveis.

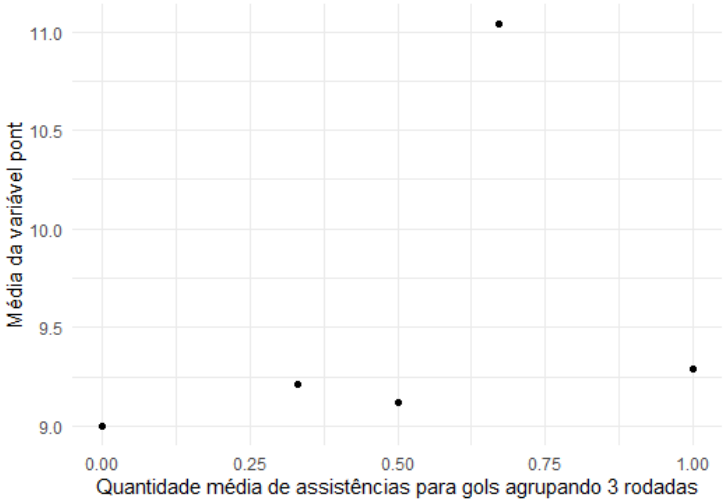

Figura 4.23: Diagrama de dispersão da quantidade agrupada de assistências para gol dadas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.23 apresenta o diagrama de dispersão da quantidade agrupada de assistências para gol dadas pelo jogador versus a média da pontuação agrupada do jogador. Analisando a mesma, podemos dizer que aparentemente não há uma tendência nos pontos, visto que em 4 dos 5 pontos praticamente não há oscilação no valor médio da variável reposta, o que evidencia a inexistência de uma possível correlação entre as duas variáveis.

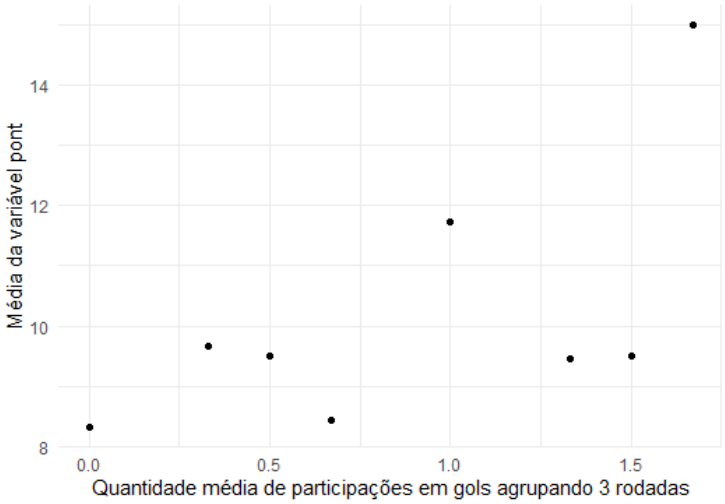

Figura 4.24: Diagrama de dispersão da quantidade agrupada de participações em gols do jogador versus a média da pontuação agrupada do jogador.

A Figura 4.24 traz o diagrama de dispersão da quantidade agrupada de participações em gols do jogador versus a média da pontuação agrupada do jogador. Por meio dela, vemos que aparentemente não há correlação linear entre as variáveis, visto que os pontos apresentam oscilações de aumentos e diminuições na média da variável resposta conforme aumentamos a variável preditora.

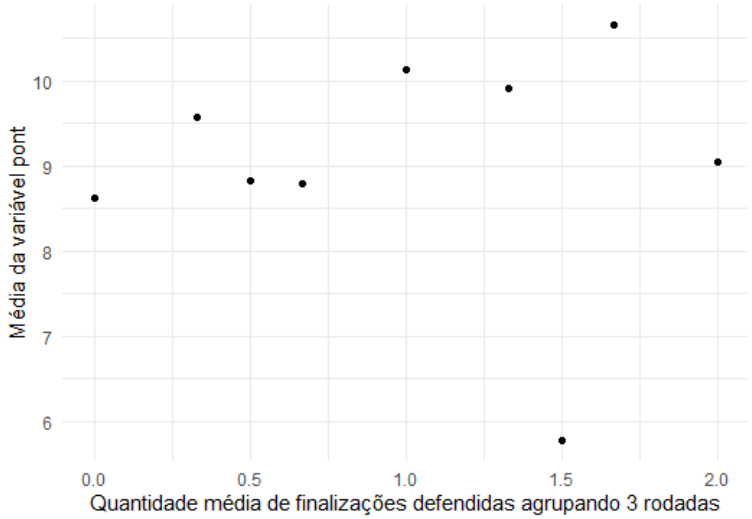

Figura 4.25: Diagrama de dispersão da quantidade agrupada de finalizações defendidas pelo goleiro adversário dadas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.25 traz o diagrama de dispersão da quantidade agrupada de finalizações defendidas pelo goleiro adversário dadas pelo jogador versus a média da pontuação agrupada do jogador. Dessa forma, vemos que há uma baixa oscilação na maior parte dos pontos, o que é um indício de que não deve haver uma correlação entre as duas variáveis.

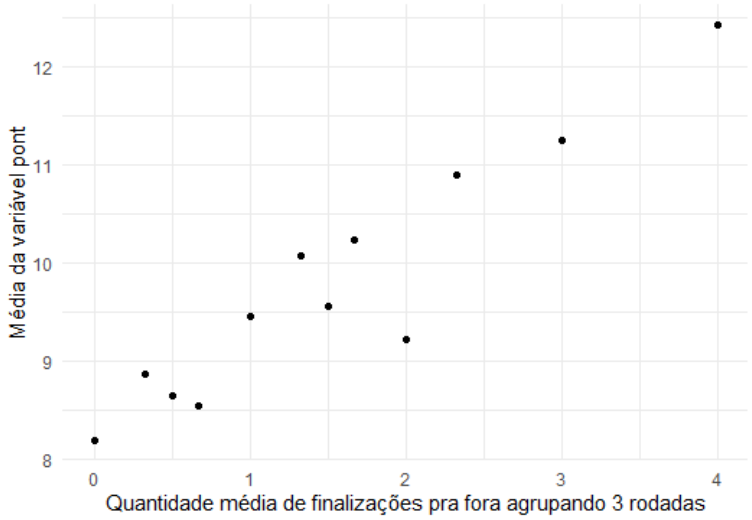

Figura 4.26: Diagrama de dispersão da quantidade agrupada de finalizações pra fora dadas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.26 apresenta o diagrama de dispersão da quantidade agrupada de finalizações pra fora dadas pelo jogador versus a média da pontuação agrupada do jogador. Da mesma, observamos que é possível enxergar uma tendência crescente nos pontos, o que permite dizer que, aparentemente, existe correlação entre as duas variáveis.

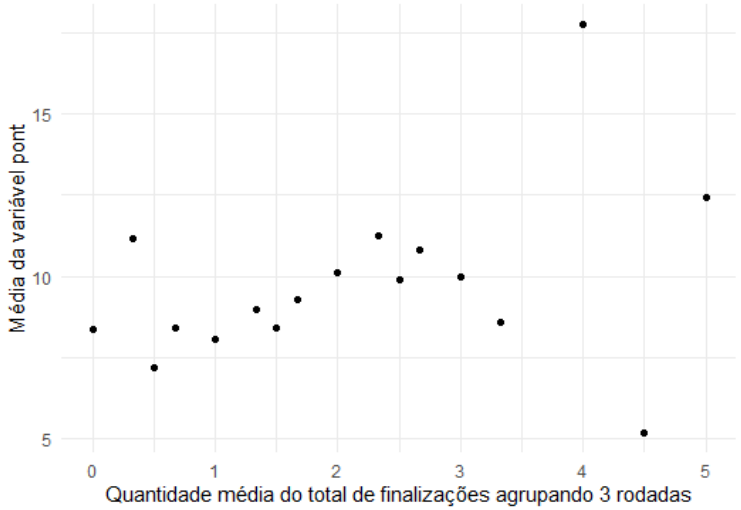

Figura 4.27: Diagrama de dispersão da quantidade agrupada do total de finalizações dadas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.27 apresenta o diagrama de dispersão da quantidade agrupada do total de finalizações dadas pelo jogador versus a média da pontuação agrupada do jogador. Por ela, observamos que, de maneira geral, há uma tendência crescente evidente nos pontos, a exceção fica por conta de alguns poucos valores mais extremos da variável preditora. No entanto, parece haver a presença de correlação entre as duas variáveis.

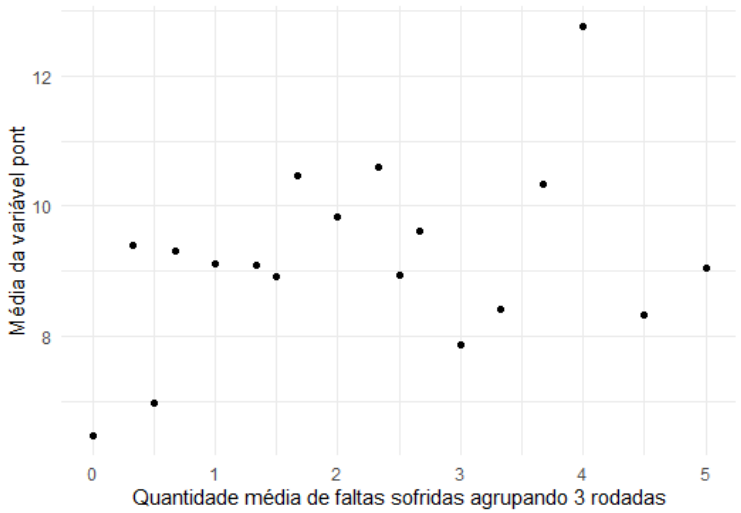

Figura 4.28: Diagrama de dispersão da quantidade agrupada de faltas sofridas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.28 traz o diagrama de dispersão da quantidade agrupada de faltas sofridas pelo jogador versus a média da pontuação agrupada do jogador. Podemos perceber que não parece existir um padrão evidente nos pontos, o que nos sugere que não há correlação entre as variáveis.

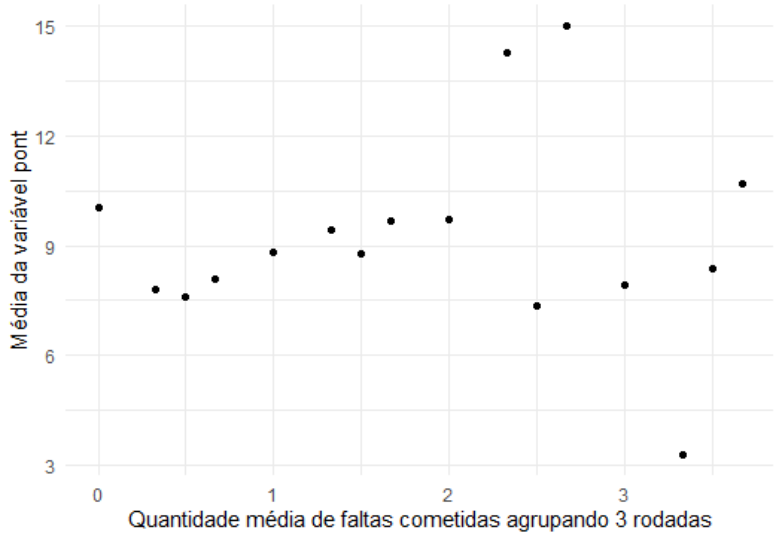

Figura 4.29: Diagrama de dispersão da quantidade agrupada de faltas cometidas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.29 apresenta o diagrama de dispersão da quantidade agrupada de faltas cometidas pelo jogador versus a média da pontuação agrupada do jogador. A partir dela, observamos que, de maneira geral, há uma oscilação nos pontos, sem que haja uma clara tendência, o que sugere considerarmos que não existe correlação entre as duas variáveis.

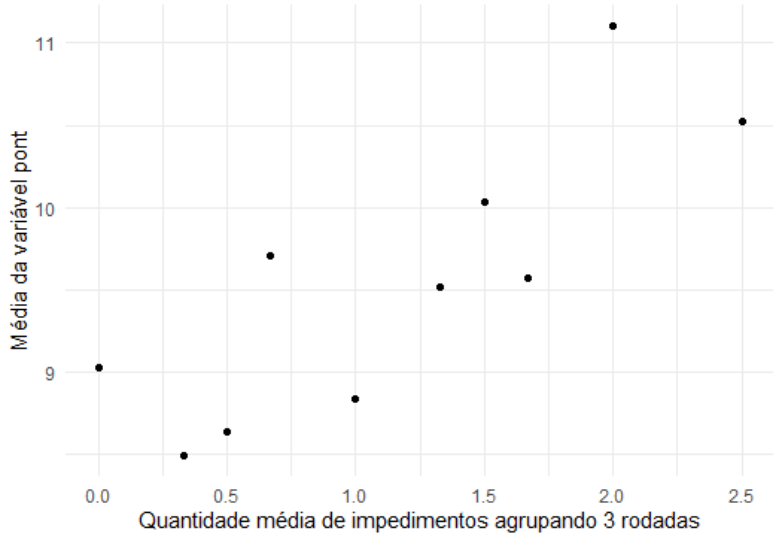

Figura 4.30: Diagrama de dispersão da quantidade agrupada de impedimentos do jogador versus a média da pontuação agrupada do jogador.

A Figura 4.30 traz o diagrama de dispersão da quantidade agrupada de impedimentos do jogador versus a média da pontuação agrupada do jogador. Da mesma, analisamos que, aparentemente, há uma certa tendência crescente nos pontos, o que é um indício da presença de correlação entre as duas variáveis.

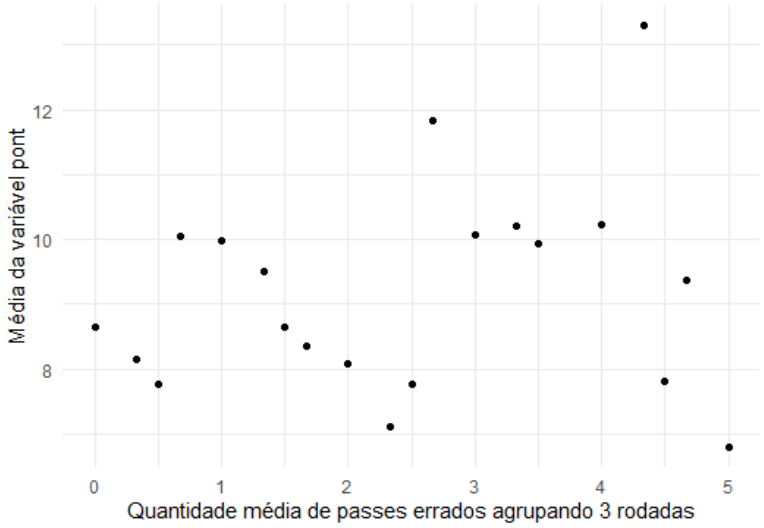

Figura 4.31: Diagrama de dispersão da quantidade agrupada de passes errados dados pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.31 apresenta o diagrama de dispersão da quantidade agrupada de passes errados dados pelo jogador versus a média da pontuação agrupada do jogador. A partir dela, percebemos que aparentemente há associação entre as variáveis, porém não linear.

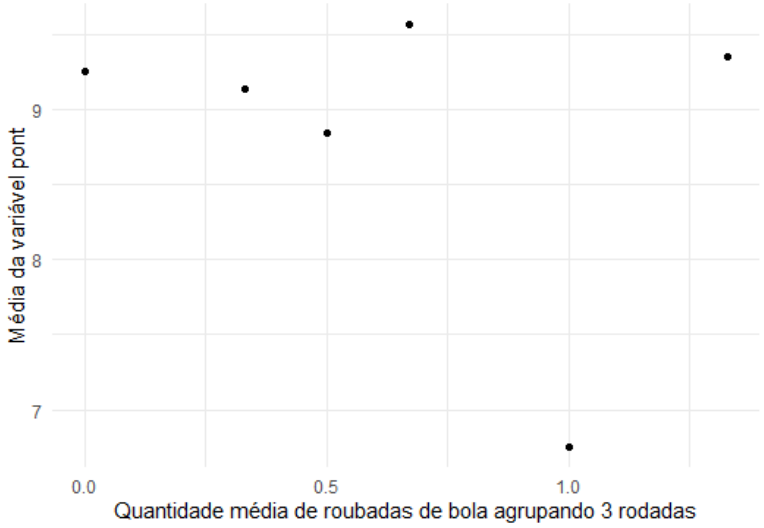

Figura 4.32: Diagrama de dispersão da quantidade agrupada de roubadas de bola feitas pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.32 traz o diagrama de dispersão da quantidade agrupada de roubadas de bola feitas pelo jogador versus a média da pontuação agrupada do jogador. A partir da mesma, vemos que em 5 dos 6 valores assumidos pela variável preditora, a média do valor da variável resposta praticamente não oscila, o que é um indício da inexistência de correlação entre as duas variáveis.

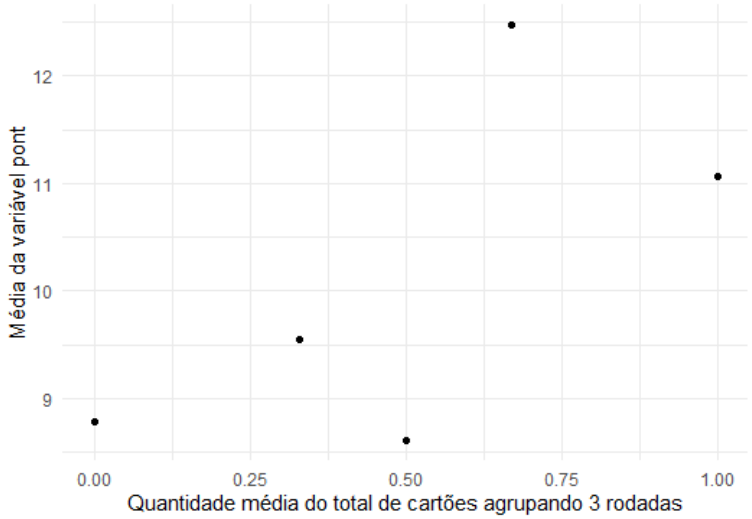

Figura 4.33: Diagrama de dispersão da quantidade agrupada do total de cartões recebidos pelo jogador versus a média da pontuação agrupada do jogador.

A Figura 4.33 traz o diagrama de dispersão da quantidade agrupada do total de cartões recebidos pelo jogador versus a média da pontuação agrupada do jogador. Apesar de não parecer haver uma associação forte entre as variáveis, os dois pontos com maior valor no eixo x também são aqueles com maior valor no eixo y.

4.2.4 Matriz de correlação

A Tabela 4.7 apresenta a matriz de correlação de Pearson considerando todas as variáveis, ou seja, nos mostra todos os coeficientes de correlação entre as variáveis. Visando facilitar a observação dessas correlações, foi feito, na Figura 4.34, um *cor plot*, que é um gráfico que ilustra as correlações das variáveis conforme a tonalidade e tamanho do círculo referente a cada combinação, ou seja, quanto mais escuro e maior o círculo, maior é, em módulo, a correlação entre as variáveis. Dessa maneira, observamos que a maioria das correlações referentes à variável resposta com alguma variável preditora são positivas, as exceções se dão com as variáveis mando de campo agrupado de partidas disputadas pelo jogador (**local**) e quantidade agrupada de roubadas de bola feitas pelo jogador (**rb**), que apresentam correlações negativas, ainda que próximas a zero. Vale ressaltar que, em alguns casos, isso ocorre até mesmo para as variáveis que, em princípio, valores mais altos indicam uma pior performance. É importante ainda dizer que a maior correlação da variável resposta é com a variável gols marcados pelo jogador (**gol**), onde o coeficiente de correlação é de 0.256. No entanto, as variáveis pontuação agrupada do jogador nas últimas 3 rodadas que ele disputou (**pont_ult**), quantidade agrupada de participações em

gols do jogador (**part**), quantidade agrupada de finalizações pra fora dadas pelo jogador (**ff**) e quantidade agrupada do total de finalizações dadas pelo jogador (**ftotal**) também apresentam um coeficiente de correlação considerável com a variável resposta, todos acima de 0.2. Esse fato, aliás, é compatível com as análises dos diagramas de dispersão feitas na Subseção 4.2.3, exceto pela variável **part**. Nesse contexto, é esperado que essas variáveis preditoras que aparentaram ter uma correlação considerável com a variável resposta nas duas análises feitas sejam significantes ou, em outras palavras, influenciem de alguma forma no estudo da variável de interesse quando for feita a modelagem dos dados.

Tabela 4.7: Matriz de correlação referente a todas as variáveis.

	pont	pont_ult	local	gol	ass	part	fd	ff	ftotal	fs	fc	i	pe	rb	cart
pont	1.000	0.255	-0.013	0.256	0.075	0.229	0.067	0.216	0.205	0.190	0.081	0.117	0.072	-0.059	0.162
pont_ult	0.255	1.000	0.093	0.679	0.522	0.813	0.208	0.298	0.345	0.237	-0.141	0.434	-0.030	0.017	-0.252
local	-0.013	0.093	1.000	0.054	0.122	0.116	0.044	-0.036	-0.004	0.038	0.088	-0.005	-0.039	-0.060	-0.011
gol	0.256	0.679	0.054	1.000	0.103	0.775	-0.024	-0.022	-0.030	0.025	0.084	0.064	-0.009	0.171	0.036
ass	0.075	0.522	0.122	0.103	1.000	0.708	-0.089	0.089	0.020	0.213	-0.015	0.146	-0.014	-0.140	-0.111
part	0.229	0.813	0.116	0.775	0.708	1.000	-0.073	0.040	-0.008	0.153	0.049	0.138	-0.015	0.031	-0.045
fd	0.067	0.208	0.044	-0.024	-0.089	-0.073	1.000	0.115	0.636	0.147	-0.116	0.045	0.163	0.122	0.077
ff	0.216	0.298	-0.036	-0.022	0.089	0.040	0.115	1.000	0.839	-0.037	-0.007	0.235	0.031	0.026	0.048
ftotal	0.205	0.345	-0.004	-0.030	0.020	-0.008	0.636	0.839	1.000	0.051	-0.069	0.207	0.114	0.087	0.079
fs	0.190	0.237	0.038	0.025	0.213	0.153	0.147	-0.037	0.051	1.000	0.086	0.017	0.285	0.044	-0.053
fc	0.081	-0.141	0.088	0.084	-0.015	0.049	-0.116	-0.007	-0.069	0.086	1.000	0.117	0.065	0.107	0.319
i	0.117	0.434	-0.005	0.064	0.146	0.138	0.045	0.235	0.207	0.0177	0.117	1.000	0.042	-0.009	0.000
pe	0.072	-0.030	-0.039	-0.009	-0.014	-0.015	0.163	0.031	0.114	0.285	0.065	0.042	1.000	-0.097	0.043
rb	-0.059	0.017	-0.060	0.171	-0.140	0.031	0.122	0.026	0.087	0.044	0.107	-0.009	-0.097	1.000	0.051
cart	0.162	-0.252	-0.011	0.036	-0.111	-0.045	0.077	0.048	0.079	-0.053	0.319	0.000	0.043	0.051	1.000

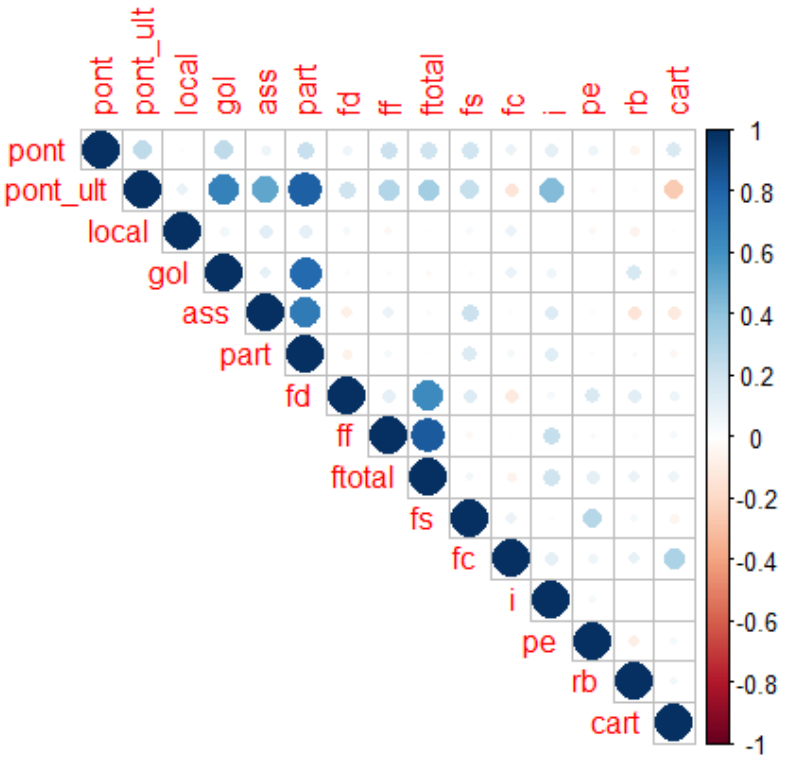

Figura 4.34: *Cor plot* referente a todas as variáveis.

Pode-se notar ainda que há algumas correlação consideráveis entre variáveis preditoras, como entre **pont_ult** e **part**, **pont_ult** e **gol** e, **pont_ult** e **ass**, cujos coeficientes de correlação são, respectivamente, 0.813, 0.679 e 0.522, o que pode ser explicado pelo fato de que no Cartola FC esses *scouts* são os de maiores pontuações, como visto no Capítulo 1. Outras grandes correlações, como é de se esperar, se dão entre as variáveis que foram criadas a partir de outras com as suas respectivas derivações, como a variável **part**, que é a soma das variáveis **gol** e **ass** e a variável **ftotal**, que é a soma das variáveis **fd** e **ff**. Dessa forma, temos que os coeficientes de correlação entre **part** e **gol**, **part** e **ass**, **ftotal** e **fd** e, **ftotal** e **ff** são, respectivamente, 0.775, 0.708, 0.636 e 0.839. No entanto, quando for feito o ajuste do modelo será importante evitar que duas variáveis preditoras fortemente correlacionadas entrem simultaneamente no modelo para evitar problemas de multicolinearidade (Draper e Smith, 1998). De acordo com Miloca e Conejo (2013), o que precisa ser feito é procurar variáveis preditoras que tenham baixa multicolinearidade com as outras variáveis preditoras, mas também apresentem correlações consideráveis com a variável resposta.

4.3 Modelagem

Ao longo desta seção será discutido como foi feita a modelagem dos dados nesse trabalho. Na Subseção 4.3.1 será descrito o pacote que foi utilizado no R para ajustar modelos via EEGs. A seguir, na Subseção 4.3.2, apresentamos os critérios usados para ajustarmos os modelos. As medidas escolhidas para avaliar o poder preditivos dos modelos e seus respectivos valores estão na Subseção 4.3.3. Por fim, considerações finais são expostas na Subseção 4.3.4.

4.3.1 Descrição do pacote *geepack* do R

A biblioteca *geepack* é uma das mais utilizadas quando deseja-se criar modelos estatísticos via EEGs. Primeiramente, para instalar o pacote, devemos digitar o seguinte comando no *software* R:

```
install.packages("geepack")
```

Após a instalação devemos carregar o pacote no R com o comando:

```
library("geepack")
```

Para ajustar modelos via EEGs com o pacote, é preciso executar a seguinte função:

```
    geeglm(
formula,
id = ,
data =  ,
family = ,
corstr = ,
waves =
)
```

cujos argumentos do pacote e suas respectivas descrições podem ser vistas na Tabela 4.8:

Tabela 4.8: Descrição dos argumentos da função *geeglm* do pacote *geepack*.

Argumento	**Descrição**
formula	A expressão de fórmula é da forma: variável resposta $\sim$ variáveis preditoras.
id	Um vetor que identifica os i indivíduos. O comprimento do argumento id deve ser o mesmo que o número de observações.
data	Argumento que especifica o banco de dados em que estão inseridas todas as observações, contendo as variáveis escolhidas nos argumentos de formula, id e waves.
family	Nesse argumento escolhemos a distribuição para a variável resposta e sua respectiva função de ligação.
corstr	Argumento que define a estrutura de correlação dos dados. As possibilidades são: "independence", "exchangeable", "ar1", "unstructured"e "userdefined".
waves	Argumento que especifica as j repetições das medidas de cada indivíduo ao longo do tempo.

Um exemplo de código aplicado no R utilizando o banco de dados usado no trabalho é dado por:

```
    ajuste_exemplo = geeglm(
pont ~ gol + ff,
id = id,
data = banco.novo,
family = Gamma (link = "log"),
corstr = "ar1",
waves = rod
)
```

Nesse exemplo, podemos observar que consideramos a variável **pont** como resposta e as variáveis **gol** e **ff** como preditoras. Além disso, a variável **id** foi a escolhida para representar os indivíduos, visto que ela é um identificador de cada jogador. Ainda, podemos perceber que a distribuição utilizada foi a gama com função de ligação logarítmica e a estrutura de correlação escolhida foi a auto regressiva de primeira ordem. Por fim, a variável **rod** é a que especifica as repetições das medidas de cada jogador ao longo do tempo.

A biblioteca *geepack* ainda nos retorna os valores de $Q(\boldsymbol{y};\boldsymbol{\mu})$, QIC, QIC_u e CIC aplicando no ajuste a seguinte função:

```
QIC()
```

4.3.2 Ajuste dos modelos

Inicialmente, antes de ajustarmos os modelos, é importante relembrar que o banco de dados foi dividido em duas partes, sendo que a primeira delas será usada para o ajuste dos modelos e a segunda para a avaliação do poder preditivo dos mesmos, conforme foi explicado no final da Subseção 4.2.1.

Ajustaremos modelos variando as seguintes características:

- Metodologia (podendo ser MLG ou EEG);
- Distribuição (podendo ser normal, gama ou gaussiana inversa);
- Função de ligação (sendo identidade para o caso em que distribuição escolhida for normal e, logarítmica ou identidade caso seja gama ou gaussiana inversa);
- Estrutura de correlação de $\boldsymbol{R}(\boldsymbol{\alpha})$ (podendo ser AR(1) ou permutável) para quando a metodologia escolhida for EEG;
- Forma de seleção das variáveis (sendo via *stepwise* AIC ou *stepwise* p-valor para o caso em que a metodologia escolhida for MLG e, *stepwise* QIC ou *stepwise* p-valor caso seja via EEG).

Na forma de seleção das variáveis, o *stepwise* QIC será feito da mesma maneira do *stepwise* AIC, apresentado na Seção 2.5. Por outro lado, via *stepwise* p-valor, pode ser resumido em, primeiramente, ajusta-se todos os modelos possíveis considerando apenas

uma variável preditora presente e seleciona aquele em que a variável apresenta o menor p-valor, desde que esse seja menor do que 0.05. Em seguida, são ajustados todos os modelos que contenham a variável inclusa no primeiro passo mais outra que ainda não foi inserida e, é escolhido o que apresentar o menor p-valor da variável inserida, desde que esse seja menor do que 0.05. Ainda nesse passo, é verificada a possibilidade de retirar a primeira variável selecionada, visto que o seu p-valor pode ter se alterado após a inclusão da nova variável, utilizaremos como critério de permanência no modelo um p-valor menor do que 0.1. Dessa maneira, repete-se esses passos de avaliar a inclusão e exclusão das variáveis até que não tenha nenhuma variável nova para ser inserida ou excluída, tendo em vista o critério de p-valor menor do que 0.05 para a inserção e o de p-valor menor do que 0.1 para a permanência no modelo.

O motivo de considerarmos não só as EEGs para o ajuste dos modelos, mas também os MLGs, é que, apesar de, tecnicamente, a primeira técnica ser mais adequada para esse trabalho, destacamos mais uma vez que o foco principal do mesmo é preditivo.

Além dos modelos via MLGs e EEGs, ajustaremos também um modelo sem envolver inferências estatísticas, o qual chamaremos de alternativo, em que consideraremos que a variável resposta, pontuação do agrupada do jogador, será igual a variável preditora pontuação agrupada do jogador nas últimas 3 rodadas. Isto é, esperaremos que a pontuação que o jogador fará em um grupo de 3 rodadas será a mesma que ele fez no grupo anterior de 3 rodadas.

Por fim, vale ressaltar que nos ajustes dos modelos em que a distribuição considerada é a normal, as variáveis **pont** e **pont_ult** assumem os seus valores reais, sem a adição da constante de valor 5 explicada na Subseção 4.2.1.

4.3.3 Poder preditivo

Existem diversas maneiras para avaliar o poder preditivo de modelos ajustados, isto é, verificar se o ajuste está retornando boas estimações do valor da variável resposta.

Nesse trabalho utilizaremos duas medidas como critérios de avaliação, a raiz do erro quadrático médio (REQM) e o erro absoluto médio (EAM). Em ambos, quanto menor o valor obtido, melhor é o ajuste do modelo. Sendo assim, a REQM é dado por

$$REQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

em que n é o número total de observações, y_i é o valor da i-ésima observação e $\hat{y}_i$ é o valor estimado da i-ésima observação.

Apesar de diferentes, a fórmula do EAM constitui dos mesmos componentes presentes na REQM e é dada por

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.2}$$

A seguir, na Tabela 4.9, apresentaremos todos os modelos ajustados na base de dados de treino e seus respectivos poderes preditivos via REQM e EAM, que foram obtidos a partir de suas aplicações na base de dados de teste.

Tabela 4.9: Poder preditivo de todos os modelos ajustados.

Metodologia	**Distribuição**	**Função de ligação**	**Estrutura de correlação de $R(\alpha)$**	**Forma de seleção**	**Poder preditivo via REQM**	**Poder preditivo via EAM**
Alternativo	-	-	-	-	3.875	3.092
MLG	Normal	Identidade	-	*stepwise* AIC	3.885	3.051
MLG	Normal	Identidade	-	*stepwise* p-valor	3.853	3.081
MLG	Gama	Identidade	-	*stepwise* AIC	3.817	3.024
MLG	Gama	Identidade	-	*stepwise* p-valor	**3.712**	3.055
MLG	Gama	Logarítmica	-	*stepwise* AIC	**3.784**	**2.988**
MLG	Gama	Logarítmica	-	*stepwise* p-valor	**3.622**	**3.014**
MLG	G. Inversa	Identidade	-	*stepwise* AIC	**3.741**	**2.934**
MLG	G. Inversa	Identidade	-	*stepwise* p-valor	4.008	3.279
MLG	G. Inversa	Logarítmica	-	*stepwise* AIC	3.923	3.138
MLG	G. Inversa	Logarítmica	-	*stepwise* p-valor	**3.645**	3.033
EEG	Normal	Identidade	AR(1)	*stepwise* QIC	3.856	**2.974**
EEG	Normal	Identidade	AR(1)	*stepwise* p-valor	**3.713**	**2.994**
EEG	Normal	Identidade	Permutável	*stepwise* QIC	3.832	**2.973**
EEG	Normal	Identidade	Permutável	*stepwise* p-valor	4.194	3.267
EEG	Gama	Identidade	AR(1)	*stepwise* QIC	4.439	3.450
EEG	Gama	Identidade	AR(1)	*stepwise* p-valor	3.920	3.138
EEG	Gama	Identidade	Permutável	*stepwise* QIC	4.431	3.437
EEG	Gama	Identidade	Permutável	*stepwise* p-valor	4.201	3.274
EEG	Gama	Logarítmica	AR(1)	*stepwise* QIC	4.438	3.446
EEG	Gama	Logarítmica	AR(1)	*stepwise* p-valor	4.164	3.217
EEG	Gama	Logarítmica	Permutável	*stepwise* QIC	4.432	3.433
EEG	Gama	Logarítmica	Permutável	*stepwise* p-valor	4.076	3.269

Primeiramente, antes de analisar os resultados, é importante dizer que, utilizando o *geepack*, não foi possível ajustar modelos via EEGs quando a distribuição escolhida para a variável resposta é a gaussiana inversa, pois o pacote emite uma mensagem afirmando que a variância é inválida.

Dito isto, podemos perceber através da Tabela 4.9 os valores destacados em negrito, que são os 6 menores e, consequentemente, os melhores resultados da avaliação do poder preditivo via REQM e EAM dentre todos os modelos ajustados. Sendo assim, com base no critério de REQM para medir o poder preditivo, o modelo ajustado que apresenta o menor valor, de 3.622, é um MLG, em que a distribuição escolhida foi a gama com função

de ligação logarítmica e a forma de seleção via *stepwise* p-valor. Por outro lado, usando o EAM como critério para avaliar o poder preditivo, temos que o modelo ajustado que apresentou o melhor resultado, de 2.934, foi, novamente, um MLG, mas dessa vez com a gaussiana inversa como distribuição, a identidade como função de ligação e o *stepwise* AIC como forma de seleção. Ainda, vale ressaltar que ambos os modelos citados estão também entre os 6 melhores segundo a outra medida.

Além de tudo o que foi dito anteriormente, é importante destacar o fato de que, considerando as duas medidas de avaliação do poder preditivo dos modelos ajustados, há um EEG que se destaca, estando presente entre os 6 melhores nas duas métricas, o mesmo apresenta distribuição normal, com estrutura de correlação AR(1) e forma de seleção via *stepwise* p-valor. Ao comparar o valor de REQM desse modelo via EEG em relação ao melhor modelo ajustado segundo o critério de REQM, notamos que o EEG apresenta um valor 2.5% maior da medida em questão. Agora, ao comparar o valor de EAM do modelo ajustado via EEG em relação ao melhor modelo ajustado segundo o critério de EAM, vemos que o modelo ajustado via EEG apresenta aumento de 2% na métrica. Por fim, tendo em vista o modelo alternativo, que foi criado sem considerar métodos estatísticos para a sua criação, o mesmo apresenta, em relação aos valores de suas medidas para a avaliação do poder preditivo, um aumento de, aproximadamente, 7% e 5.4% ao ser comparado com os melhores modelos ajustados via REQM e EAM, respectivamente.

4.3.4 Considerações sobre os modelos de melhor poder preditivo

Como foi visto na subseção anterior, o modelo ajustado que apresentou o melhor poder preditivo utilizando o critério de REQM foi um MLG, em que a distribuição escolhida foi a gama, com função de ligação logarítmica e a forma de seleção via p-valor. Enquanto utilizando o EAM como critério, foi um MLG, com a gaussiana inversa como distribuição, a identidade como função de ligação e o *stepwise* AIC como forma de seleção. Sendo assim, apresentaremos a seguir o formato desses modelos ajustados, com suas respectivas variáveis inseridas e a estimativa do vetor de parâmetros $\boldsymbol{\beta}$ para cada um.

Começando pelo modelo ajustado escolhido utilizando o critério de EAM para a avaliação do poder preditivo, temos

- $Y_i \overset{ind}{\sim}$ Gaussiana Inversa$(\mu_i,\, \phi)$, $i = 1, ..., 123$;
- $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8}$.

em que:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \\ \hat{\beta}_7 \\ \hat{\beta}_8 \end{bmatrix} = \begin{bmatrix} 8.570 \\ -0.418 \\ 190.128 \\ 185.720 \\ -182.720 \\ 1.059 \\ 0.588 \\ 1.018 \\ -2.599 \end{bmatrix} \text{ e } \boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \\ x_{i5} \\ x_{i6} \\ x_{i7} \\ x_{i8} \end{bmatrix} = \begin{bmatrix} 1 \\ pont\_ult_i \\ gol_i \\ ass_i \\ part_i \\ ftotal_i \\ fs_i \\ i_i \\ rb_i \end{bmatrix}.$$

Podemos perceber que foi selecionada uma grande quantidade de variáveis nesse modelo. Além disso, esse modelo não aparenta ser interessante, pois apresenta estimativas de sinal contrário ao esperado pela análise descritiva, sugerindo a existência de multicolinearidade. A mesma é evidente não só por isso, mas também por algumas estimativas com valores expressivos tendo em vista as demais, que é o que ocorre nas variáveis **gol**, **ass** e **part**, em que as estimativas dos parâmetros referentes a elas apresentam valores de 190.128, 185.720 e -182.720, respectivamente, enquanto que as demais estimativas (exceto o intercepto) estão contidas no intervalo de [-2.6,1.06]. Por fim, vale ressaltar que, como vimos na Subseção 4.2.4, já haviam indícios da existência de multicolinearidade entre as variáveis citadas, visto que os altos valores de correlação das variáveis **ass** e **gol** com a **part** se deve ao fato de que, a segunda, foi criada a partir das outras duas, sendo nada mais do que a soma das mesmas.

Agora, considerando o modelo ajustado escolhido utilizando o critério de REQM para a avaliação do poder preditivo, temos

- $Y_i \overset{ind}{\sim} \text{Gama}(\mu_i, \phi)$, $i = 1, ..., 123$;
- $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$.

em que:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 1.847 \\ 0.034 \\ 0.274 \end{bmatrix} \text{ e } \boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \end{bmatrix} = \begin{bmatrix} 1 \\ pont\_ult_i \\ cart_i \end{bmatrix}.$$

Podemos ver que apenas duas variáveis foram selecionadas no ajuste desse modelo, **pont_ult** e **cart**, sendo que a última não está presente no melhor modelo via EAM, o qual mostramos anteriormente. Ainda, vale ressaltar que os sinais das estimativas estão condizentes com o sugerido pela análise descritiva, visto que ambas as variáveis incluídas no modelo apresentam correlação positiva com a resposta. É importante dizer ainda que, um dos possíveis motivos para justificar a não entrada das outras variáveis preditoras no ajuste desse modelo, é o fato de que quase todas estão diretamente relacionadas com a variável **pont_ult**, apresentando correlações consideráveis com a mesma, como podemos ver na Tabela 4.7 e na Figura 4.34.

Dentre os dois modelos apresentados anteriormente, assumiremos o escolhido via REQM como melhor, visto que não aparenta ter problemas de multicolinearidade como o escolhido via EAM. Sendo assim, como o modelo correspondente a esse considerado ideal não foi selecionado por nenhuma das formas de seleção utilizadas em EEGs, a intenção é ajustá-lo via EEG considerando as suas características, variando apenas a estrutura de correlação de $\boldsymbol{R}(\boldsymbol{\alpha})$, com o intuito de comparar o poder preditivo do modelo via MLG e via EEG.

Tabela 4.10: Poder preditivo dos modelos EEGs correspondentes ao escolhido via REQM.

Metodologia	Distribuição	Função de ligação	Estrutura de correlação de $R(\alpha)$	Poder preditivo via REQM	Poder preditivo via EAM
EEG	Gama	Logarítmica	AR(1)	3.625	2.970
EEG	Gama	Logarítmica	Permutável	3.974	3.174

A Tabela 4.10 apresenta os poderes preditivos dos modelos ajustados correspondentes ao escolhido via REQM, mas considerando as EEGs. Da mesma, podemos perceber que no caso em que a estrutura de correlação escolhida para $\boldsymbol{R}(\boldsymbol{\alpha})$ foi AR(1), o ajuste apresentou bons resultados de poder preditivo. Considerando o REQM para a avaliação, o seu valor é de 3.625 e, apesar de ser pior que o valor de 3.622 do modelo correspondente via MLG, com exceção do mesmo, ele seria o melhor dentre todos os outros ajustes que foram feitos na Tabela 4.9. Além disso, tendo em vista o EAM para a avaliação do poder preditivo, o seu valor foi de 2.97, que é melhor do que o 3.014 apresentado pelo modelo via MLG e, colocaria esse ajuste com o segundo melhor valor de EAM dentre todos os ajustes. Por fim, o ajuste do modelo considerando a estrutura de correlação permutável para $\boldsymbol{R}(\boldsymbol{\alpha})$ não apresentou bons resultados de poder preditivo em nenhum dos critérios considerados, sendo ambos piores que no modelo correspondente via MLG.

Como o modelo ajustado via EEG com estrutura de correlação AR(1) para $\boldsymbol{R}(\boldsymbol{\alpha})$ obteve o segundo melhor resultado considerando ambos os critérios de avaliação do poder preditivo e, além disso, é metodologicamente mais adequado do que o seu correspondente via MLG, escolhemos este modelo após esta etapa. No entanto, como há apenas duas variáveis no mesmo, verificaremos, a partir desse ajuste, a possibilidade de incluir mais variáveis, com o intuito de melhorar o poder preditivo. A seleção das novas variáveis consistirá em, primeiramente, ajustar todos os modelos considerando as duas variáveis já presentes, **pont_ult** e **cart**, mais outra que ainda não foi inserida e selecionar o que retorna o menor valor de REQM, desde que esse seja menor do que o modelo ajustado com as duas variáveis iniciais. Em seguida, são ajustados todos os modelos que contenham as duas variáveis de início, a inclusa no passo anterior, mais outra que ainda não foi inserida e, é escolhido o que apresentar o menor valor de REQM, desde que esse seja menor do que o modelo ajustado no passo anterior. Dessa forma, repete-se esses passos de inclusão de variáveis até que não haja diminuição no valor de REQM. O mesmo será feito, de forma equivalente, considerando o EAM como critério de avaliação do poder preditivo.

O modelo ajustado final obtido a partir da seleção de novas variáveis foi idêntico em ambas as medidas de poder preditivo avaliadas e é dado por

- $Y_{ij} \sim \text{Gama}(\mu_{ij},\, \phi)$, $i = 1, ..., 31$ e $j = 1, ..., 5$;
- $\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}$;
- $\text{Corr}(Y_{ij}, Y_{ij'}) = 1$ para $j = j'$ e $\text{Corr}(Y_{ij}, Y_{ij'}) = \rho^{|j-j'|}$ para $j \neq j'$.

em que:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 1.515 \\ 0.070 \\ 0.167 \\ 0.020 \end{bmatrix} \text{ e } \boldsymbol{x_{ij}} = \begin{bmatrix} 1 \\ x_{ij1} \\ x_{ij2} \\ x_{ij3} \end{bmatrix} = \begin{bmatrix} 1 \\ pont\_ult_{ij} \\ cart_{ij} \\ pe_{ij} \end{bmatrix}.$$

Como podemos observar, apenas uma nova variável foi inserida no ajuste do modelo, **pe**. Analisando o modelo ajustado percebemos que os sinais das estimativas dos coeficientes do vetor de parâmetros $\hat{\boldsymbol{\beta}}$ estão condizentes com o esperado a partir da análise descritiva, além de não aparentar haver problemas de multicolinearidade. Os valores de REQM e EAM do mesmo são, respectivamente, 3.566 e 2.932, o que torna esse modelo

ajustado, tendo em vista ambos os critérios de avaliação do poder preditivo, o melhor dentre todos os ajustados na Tabela 4.9, pois apresenta o menor valor nas duas métricas.

Levando em consideração o que foi visto através da análise descritiva que foi feita ao longo da Seção 4.2, era esperado que a variável **pont_ult** fosse significante para o estudo da variável resposta. Já a variável **cart** não está entre aquelas que apresentam maior correlação com a variável resposta. No entanto, entre as variáveis com correlação com a resposta superior a 0.15, ela é aquela que apresenta a menor correlação com a variável **pont_ult**. Isso provavelmente explica a sua inclusão no modelo e a não entrada de outras variáveis que apresentam maior correlação com a resposta. Por outro lado, a variável **pe** apresenta correlação amostral inferior a 0.10 e não era esperado que fosse incluída no modelo. No entanto, é interessante destacar que, em geral, a finalidade principal da análise descritiva é verificar se há problemas no conjunto de dados em estudo. Tendo em vista isso que foi dito, aliado ao fato de que o foco do trabalho está na predição, não vemos como um problema a presença da variável **pe** no modelo ajustado, dado que as estimativas dos parâmetros associadas as mesmas apresentaram valores coerentes com o esperado e aparentemente sem a existência de multicolinearidade.

É curioso que as estimativas dos parâmetros referentes às variáveis **cart** e **pe** apresentem valor positivo, já que em princípio era esperado que o número de passes errados e o de cartões apresentassem correlação negativa com a pontuação agrupada do jogador. Porém, já na análise descritiva, observou-se correlação positiva entre essas variáveis e a variável resposta.

Por fim, apresentaremos a interpretação das estimativas dos coeficientes do vetor de parâmetros $\hat{\boldsymbol{\beta}}$. Para isso, é necessário aplicar a exponencial nos valores dos mesmos, conforme será mostrado a seguir.

- $\exp(\hat{\beta}_0) = 4.549$.

  Estima-se que a média da pontuação agrupada do jogador (**pont**) seja 4.549 quando todas as variáveis preditoras assumem o valor zero.

- $\exp(\hat{\beta}_1) = 1.073$.

  Estima-se que a média da pontuação agrupada do jogador (**pont**) aumente 7.3% a cada aumento de uma unidade na pontuação agrupada do jogador nas últimas 3 rodadas (**pont_ult**), mantendo-se as demais variáveis preditoras constantes.

- $\exp(\hat{\beta}_2) = 1.182$.

  Estima-se que a média da pontuação agrupada do jogador (**pont**) aumente 18.2% a cada aumento de uma unidade na quantidade agrupada do total de cartões recebidos pelo jogador (**cart**), mantendo-se as demais variáveis preditoras constantes.

- $\exp(\hat{\beta}_3) = 1.020$

  Estima-se que a média da pontuação agrupada do jogador (**pont**) aumente 2.0% a cada aumento de uma unidade na quantidade agrupada de passes errados dados pelo jogador (**pe**), mantendo-se as demais variáveis preditoras constantes.

Capítulo 5

Conclusão

Neste trabalho estudamos duas metodologias para a modelagem de dados, os modelos lineares generalizados, que assumem a independência de suas observações e, as equações de estimação generalizadas, que consideram a existência de uma estrutura de correlação entre as observações de um mesmo indivíduo. Essas técnicas foram aplicadas em um banco de dados real, criado a partir de estatísticas do *fantasy game* Cartola FC, para que, posteriormente, fosse feita a avaliação do poder preditivo dos modelos ajustados, seguindo dois critérios, a raiz do erro quadrático médio e o erro absoluto médio.

Uma das principais conclusões que obtivemos nesse trabalho é que nem sempre o modelo mais adequado segundo a teoria é o que apresentará melhores resultados preditivos. Isso é dito, pois, segundo a teoria acreditamos que as EEGs, em detrimento aos MLGs, são mais adequadas para o banco de dados em questão, visto que ele se trata de um estudo longitudinal, em que observações de um mesmo jogador são coletas repetidas vezes ao longo do tempo, através de rodadas. Então, era esperado que as observações de um mesmo jogador fossem correlacionadas entre si. No entanto, vimos na avaliação do poder preditivo que, seguindo os critérios pré definidos para a análise da mesma, os modelos ajustados via MLGs apresentaram resultados semelhantes aos EEGs em ambas as medidas avaliativas, tendo em vista os melhores ajustes nas duas metodologias, apresentados na Subseção 4.3.4. Isso pode possivelmente ser justificado pelo fato de que a correlação entre as pontuações de um mesmo jogador em grupos de rodadas diferentes não era tão alta. Sendo assim, é de suma importância que o pesquisador tenha um objetivo principal bem definido desde o início do estudo. Usando esse trabalho como exemplo, em que o foco era na predição da pontuação dos atacantes no Cartola FC, vimos que, apesar de termos escolhido um ajuste via EEG como o melhor dentre todos, alguns MLGs

obtiveram resultados semelhantes ao mesmo em relação ao poder preditivo.

Para trabalhos futuros, além dos atacantes, podem ser exploradas outras posições dos jogadores do Cartola FC, como goleiros, laterais, zagueiros ou meias. Como foi dito no Capítulo 1, algumas dessas posições apresentam estatísticas específicas, como é o caso dos goleiros e as defesas difíceis (DD). Então, pode ser que haja menor variação nas pontuações dos mesmos em relação a dos atacantes, além de talvez as suas variáveis preditoras apresentarem correlações maiores com a variável resposta. Um outro trabalho possível envolve a comparação entre o poder preditivo do MLG e EEG em casos em que as observações são longitudinais. Podem ser avaliados diversos bancos de dados para estudar se é comum que, em relação ao poder preditivo, não haja perda significativa em ignorar a correlação entre as observações. Ainda, para que pudesse trabalhar com uma distribuição assimétrica, mas não precisasse adicionar uma constante à pontuação dos jogadores, poderia ser considerada a distribuição Normal Assimétrica (Azzalini, 2005). Por fim, ao invés de EGGs, uma possibilidade seria a utilização da metodologia de Modelos Mistos, cujos detalhes, para o leitor que se interessar, podem ser consultados em Singer e Andrade (1986).

Referências Bibliográficas

Agranonik, M. (2009), Equações de estimação generalizadas (gee): aplicação em estudo sobre mortalidade neonatal em gemelares de porto alegre, rs (1995-2007), Dissertação de mestrado em epidemiologia, Universidade Federal do Rio Grande do Sul.

Akaike, H. (1974), A new look at the statistical model identification, *in* 'Selected Papers of Hirotugu Akaike', Springer, pp. 215–222.

Azzalini, A. (2005), 'The skew-normal distribution and related multivariate families', *Scandinavian Journal of Statistics* **32**(2), 159–188.

Baia, L. L. (1997), As equações de estimação generalizadas e aplicações, Dissertação de mestrado em estatística, Universidade Estadual de Campinas.

Ballinger, G. A. (2004), 'Using generalized estimating equations for longitudinal data analysis', *Organizational research methods* **7**(2), 127–150.

Bové, D. S., Held, L. et al. (2011), 'Hyper-g priors for generalized linear models', *Bayesian Analysis* **6**(3), 387–410.

Brolo, C. L. (2019), Comparação da performance do lasso e do método da máxima verossimilhança com seleção de variáveis em modelos de regressão para dados binários, Trabalho de conclusão de curso, Universidade Federal de São Carlos.

Burnham, K. P. e Anderson, D. R. (2004), 'Multimodel inference: understanding aic and bic in model selection', *Sociological methods & research* **33**(2), 261–304.

CartolaFC (2019), Disponível em: < https://assine.globo.com/panfleto/globo.com-termosepoliticascartolafc.html >. Acessado em: 3 abr. 2020.

CBF (2019), Disponível em: < https://conteudo.cbf.com.br/cdn/201902/20190226183451_971.pdf >. Acessado em: 3 abr. 2020.

Cordeiro, G. M. e McCullagh, P. (1991), 'Bias correction in generalized linear models', *Journal of the Royal Statistical Society: Series B (Methodological)* **53**(3), 629–643.

Cui, J. (2007), 'Qic program and model selection in gee analyses', *The Stata Journal* **7**(2), 209–220.

Dal Bello, L. H. A. (2010), Modelagem em experimentos mistura-processo para otimização de processos industriais, PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.

Demétrio, C. G. B. (2001), *Modelos lineares generalizados em experimentação agronômica*, USP/ESALQ.

Dias, S., Sutton, A. J., Ades, A. e Welton, N. J. (2013), 'Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials', *Medical Decision Making* **33**(5), 607–617.

Draper, N. R. e Smith, H. (1998), *Applied regression analysis*, Vol. 326, John Wiley & Sons.

FIFA (2000), Disponível em: < https://www.fifa.com/news/pele-the-greatest-them-all-1656982 >. Acessado em: 3 abr. 2020.

Fitzmaurice, G. M. (1995), 'A caveat concerning independence estimating equations with multivariate binary data', *Biometrics* pp. 309–317.

GloboEsporte (2019), Disponível em: < https://globoesporte.globo.com/cartola-fc/tutoriais/noticia/duvidas-gerais-do-cartola-fc-tutorial-explica-situacoes-especificas-do-game.ghtml >. Acessado em: 3 abr. 2020.

Guo, X., Pan, W., Connett, J. E., Hannan, P. J. e French, S. A. (2005), 'Small-sample performance of the robust score test and its modifications in generalized estimating equations', *Statistics in medicine* **24**(22), 3479–3495.

Hardin, J. W. e Hilbe, J. M. (2013), *Generalized estimating equations*, 2nd edn, CRC Press.

Hin, L.-Y. e Wang, Y.-G. (2009), 'Working-correlation-structure identification in generalized estimating equations', *Statistics in medicine* **28**(4), 642–658.

Lara, I., Spyrides, M., Guerra, M. G. e Rangel, A. (2012), 'Análise comparativa de modelos para dados longitudinais no estudo da contagem do número de bactérias presentes no leite de vaca', *Revista Brasileira de Biometria* **30**, 492–508.

Liang, K. Y. e Zeger, S. L. (1986), *Longitudinal data analysis using generalized linear models*, Biometrika.

McCullagh, P. (1983), 'Quasi-likelihood functions', *The Annals of Statistics* pp. 59–67.

Miloca, S. A. e Conejo, P. D. (2013), 'Multicolinearidade em modelos de regressão', *Semana acadêmica da matemática* **22**.

Nelder, J. A. e Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3).

Neter, J., Kutner, M. H., Nachtsheim, C. J. e Wasserman, W. (1996), *Applied linear statistical models*, Vol. 4, Irwin Chicago.

Oesselmann, C. C. (2016), Equações de estimação generalizadas com resposta binomial negativa: modelando dados correlacionados de contagem com sobredispersão, PhD thesis, Universidade de São Paulo.

Pan, W. (2001), 'Akaike's information criterion in generalized estimating equations', *Biometrics* **57**(1), 120–125.

Paula, G. A. (2004), *Modelos de regressão: com apoio computacional*, IME-USP São Paulo.

Pedroso, F. M. d. T. (2007), Uma proposta para análise de dados com correlação espacial e temporal, Dissertação de mestrado em estatística, Universidade Federal de São Carlos.

Pereira, G. H. d. A. (2019), Modelos lineares generalizados, Notas de aula, Universidade Federal de São Carlos.

Rotnitzky, A. e Jewell, N. P. (1990), 'Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data', *Biometrika* **77**(3), 485–497.

Singer, J. M. e Andrade, D. d. (1986), 'Análise de dados longitudinais', *Simpósio Nacional de Probabilidade e Estatística* **7**.

Turkman, M. A. A. e Silva, G. L. (2000), Modelos lineares generalizados da teoria a prática, *in* 'VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa'.

Venezuela, K. M., Botter, D. A. e Sandoval, M. C. (2007), 'Diagnostic techniques in generalized estimating equations', *Journal of Statistical Computation and Simulation* **77**(10), 879–888.

Venezuela, M. K. (2003), Modelos lineares generalizados para análise de dados com medidas repetidas, PhD thesis, Universidade de São Paulo.

Vieira, A. (2004), Análise da média e dispersão em experimentos fatoriais não replicados para otimização de processos industriais, PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.

Wedderburn, R. W. (1974), 'Quasi-likelihood functions, generalized linear models, and the gauss—newton method', *Biometrika* **61**(3), 439–447.

Ziegler, A. (2011), *Generalized Estimating Equations*, Lecture Notes in Statistics 204, 1 edn, Springer-Verlag New York.

Apêndice A

Banco de dados novo

Os 31 atacantes que fazem parte do banco de dados novo, numerados de acordo com os seus respectivos **id** são:

1. Gabriel, do Flamengo;
2. Bruno Henrique, do Flamengo;
3. Dudu, do Palmeiras;
4. Everton, do Grêmio;
5. Paolo Guerrero, do Internacional;
6. Eduardo Sasha, do Santos;
7. Gilberto, do Bahia;
8. Everaldo, da Chapecoense;
9. Antony, do São Paulo;
10. Ricardo Bueno, do CSA;
11. Rony, do Athletico Paranaense;
12. Yony Gonzáles, do Fluminense;
13. Marrony, do Vasco;
14. Wellington Paulista, do Fortaleza;

15. Clayson, do Corinthians;
16. Michael, do Goiás;
17. Artur, do Bahia;
18. Marinho, do Santos;
19. Osvaldo, do Fortaleza;
20. Leandro Barcia, do Goiás;
21. Di Santo, do Atlético Mineiro;
22. Diego Souza, do Botafogo;
23. Fred, do Cruzeiro;
24. Pepê, do Grêmio;
25. Romarinho, do Fortaleza;
26. Leandro Carvalho, do Ceará;
27. Diego Tardelli, do Grêmio;
28. Marcelo Cirino, do Athletico Paranaense;
29. David, do Cruzeiro;
30. Nico López, do Internacional; e
31. Jonathan, do Avaí.

A seguir veremos uma amostra do banco de dados novo, onde o valor 1 da variável **rod** representa a agregação das rodadas 23 a 25 na variável resposta, o valor 2 representa a agregação das rodadas 26 a 28, o valor 3 representa a agregação das rodadas 29 a 31, o valor 4 representa a agregação das rodadas 32 a 35 e, por fim, o valor 5 representa a agregação das rodadas 36 a 38.

Tabela A.1: Banco de dados novo.

id	rod	pont	pont_ult	local	gol	ass	part	fd	ff	ftotal	fs	fc	i	pe	rb	cart
1	2	10.36	10.7	66.7	0.67	0	0.67	1	1.33	2.33	0.67	1.67	0	1.33	0.67	0.33
1	3	11.2	10.36	0	0.33	0	0.33	2	1	3	1	0.67	1	1.67	0.67	0.67
1	4	17.75	11.2	50	0.5	0	0.5	1	3	4	2	1	0	3	0	0.5
1	5	23.7	17.75	50	1	1	2	2.5	0.5	3	1	0	0	3	1.5	0.5
2	1	16.5	9.56	33.3	0.33	0	0.33	0.67	1.33	2	1.33	2.33	1.33	2.67	0	0.33
2	2	14.2	16.5	100	1	0	1	1.33	0.67	2	1.33	2.33	1.67	1	1.33	0.33
2	3	16.06	14.2	33.3	0.5	0	0.5	0.5	1.5	2	2.5	1.5	2.5	2	0.5	0
2	4	15	16.06	100	1	0.67	1.67	0.33	1.33	1.67	4	2.67	0.33	3	0	0.33
2	5	19.15	15	50	1	0	1	0	0	0	5	3	2	0	0	1
3	1	9.2	12.6	66.7	0	1	1	0.5	1	1.5	5	2	0.5	3.5	0	0
3	2	9.36	9.2	33.3	0.33	0	0.33	1.67	0.33	2	2.67	0.67	0	4.67	0.33	0.33
3	3	10.46	9.36	66.7	0	0.33	0.33	0.33	1	1.33	3.33	2	0.33	1.67	0.33	0
3	4	8.13	10.46	66.7	0	0.33	0.33	0.67	1	1.67	3	1	0.33	1	0	0
3	5	13.2	8.13	66.7	0	0	0	1.33	1.33	2.67	4	1	0.67	4	0.33	0.33
4	1	10.3	17.75	100	0.5	1	1.5	0.5	2	2.5	0.5	2	2.5	0.5	0	0.5
4	2	12.43	10.3	66.7	0	0	0	1	4	5	1	0	2	0	0	1
4	3	13.9	12.43	66.7	0.33	0	0.33	1.67	1	2.67	1.67	0	1.33	0.67	0.33	0
4	4	10.9	13.9	33.3	0.33	0.33	0.67	1	2.33	3.33	1.33	1.33	0.67	0	0	0
4	5	9.36	10.9	66.7	0.33	0	0.33	0.33	1.33	1.67	2	0.33	1	1	0.33	0
5	1	6.7	4.3	0	0	0	0	1	1.5	2.5	1.5	0.5	0	2	1	0.5
5	2	12.76	6.7	33.3	0	0	0	0	3	3	1	0	0	4	0	0
5	3	7.76	12.76	33.3	0.67	0	0.67	1	1.67	2.67	1.33	2	0.67	2	0.33	0
5	4	11.8	7.76	66.7	0	0.33	0.33	0	0.67	0.67	4	1	0.67	2	0	0
5	5	12.2	11.8	33.3	0.67	0	0.67	1.33	1	2.33	2	2	0	0.33	0.33	0.33
6	1	7.63	9.16	33.3	0.33	0	0.33	0.33	0.67	1	1	0.33	0.67	2	0.33	0
6	2	7.93	7.63	33.3	0	0	0	0.33	1.33	1.67	1.67	0.67	1	1.67	0.67	0
6	3	15.66	7.93	66.7	0.33	0	0.33	0.33	0	0.33	0.67	0	0	1	0.33	0
6	4	9.46	15.66	66.7	0.67	0.67	1.33	0.33	0.33	0.67	1.33	0.33	1	1.67	0.33	0
6	5	4.77	9.46	33.3	0.33	0	0.33	1.67	0.33	2	0.67	0.33	0	1	0.67	0
7	1	15.1	10.93	50	0.33	0.33	0.67	0	1	1	0.33	1.33	1	0.67	0	0
7	2	5.2	15.1	66.7	0	0	0	1.5	3	4.5	0	0.5	2.5	2	0	0
7	3	7.9	5.2	50	0	0	0	0.33	1.33	1.67	0.33	2	0.33	0.33	0.33	0.33
7	4	9.86	7.9	33.3	0	0	0	2	0.5	2.5	1	0	0	0.5	0.5	0
7	5	12.5	9.86	66.7	0.33	0	0.33	0.33	1	1.33	1	1	1	1.67	0	0
8	1	6.75	3.75	50	0	0	0	0	0.5	0.5	1	3.5	0.5	0.5	0	0.5
8	2	9.1	6.75	33.3	0	0	0	0.5	1.5	2	2.5	1.5	1	3.5	0	0.5
8	3	9.46	9.1	33.3	0.67	0	0.67	0.33	1.33	1.67	1	3.67	0.33	4	0	0.33
8	4	12.76	9.46	66.7	0.33	0	0.33	0.33	1	1.33	1	1	0	1.33	0	0
8	5	7.3	12.76	100	0.67	0	0.67	0.67	2.33	3	1.67	2	0.67	1.33	0	0.33
9	1	14	7.16	100	0	0	0	1	0.67	1.67	2	1.33	0.67	2.67	0	0
9	2	8.36	14	66.7	0	1	1	1	2	3	4	1	0	1	0	0
9	3	13.3	8.36	33.3	0	0.33	0.33	0.33	0.33	0.67	1.67	1.67	2	4.33	0	0.33
9	4	6.8	13.3	50	0.33	0.33	0.67	1.33	0.67	2	3.33	0.67	1.33	5	0	0
9	5	15.16	6.8	66.7	0	0	0	1.5	0.5	2	3	1	0.5	3	0.5	0.5
10	1	9.7	9.33	100	0.33	0	0.33	0	0.33	0.33	2.33	1	0.67	0.33	0.33	0

Continua na próxima página

Tabela A.1: Continuação

id	rod	pont	pont_ult	local	gol	ass	part	fd	ff	ftotal	fs	fc	i	pe	rb	cart
10	2	11.45	9.7	0	0.5	0	0.5	0.5	0.5	1	1	0	0.5	1	0.5	0.5
10	3	8.33	11.45	33.3	0.5	0	0.5	0.5	0	0.5	4.5	1	0.5	0.5	1	0
10	4	7.96	8.33	66.7	0.33	0	0.33	0.67	0.67	1.33	3.33	1	0	1.67	0	0.67
10	5	8.36	7.96	33.3	0	0	0	0.33	1.33	1.67	2.67	1.67	0.33	1.67	0	0
11	1	9.3	8.4	33.3	0	0.33	0.33	0.33	3	3.33	1	1.67	0.67	1.33	1.33	0.33
11	2	15.13	9.3	66.7	0	0	0	1	1.33	2.33	3.67	2	2	2.67	0.67	0.33
11	3	8.7	15.13	66.7	0.33	0.67	1	1.67	1.33	3	0.67	1.33	1.33	2	0	0
11	4	8.4	8.7	50	0.33	0	0.33	0.67	1.33	2	2	2	0	1.67	0.67	0
11	5	10.26	8.4	66.7	0	0.5	0.5	0.5	0.5	1	2.5	3	1	2	1.5	0
12	1	9.73	10.46	33.3	0	0.67	0.67	0.33	0.33	0.67	1.33	1	1	0.67	0	0
12	2	11.5	9.73	66.7	0.33	0	0.33	1	1	2	2.33	3	0.33	0.67	0.33	0
12	3	5.53	11.5	33.3	0	0	0	1.33	2	3.33	3.67	1.67	1.33	2.33	0	0
12	4	11.86	5.53	33.3	0	0	0	0.33	1	1.33	1	2	0.67	1.67	0	0.33
12	5	6.7	11.86	66.7	0.33	0	0.33	1.33	0.33	1.67	1.33	0.67	1	1.67	0.67	0
13	1	6.5	6.65	100	0	0	0	0.5	1	1.5	1	0.5	0	0	0	0
13	2	11.46	6.5	33.3	0	0	0	0.5	1.5	2	2	1.5	0.5	1	0	0.5
13	3	6.83	11.46	66.7	0.33	0.33	0.67	0.33	0	0.33	1.67	2	1.33	0.33	0	0
13	4	13.1	6.83	100	0	0	0	0.33	1.67	2	1.67	1.67	1	1.33	0	1
13	5	11.76	13.1	33.3	0	1	1	0	2	2	5	2	0	0	0	0
14	1	9.76	11.6	33.3	0.5	0	0.5	0.5	1	1.5	1	2	0.5	1	0.5	0
14	2	17.1	9.76	50	1	0	1	0.33	0	0.33	1	2	0	2.67	0.67	1
14	3	8.7	17.1	50	1	0.5	1.5	0	1	1	2.5	1	0.5	1.5	0.5	0
14	4	10.43	8.7	66.7	0.5	0	0.5	0	1.5	1.5	1	1.5	0	0	0.5	0.5
14	5	7.63	10.43	33.3	0.33	0.33	0.67	0.33	0.33	0.67	1	1.33	0.67	1.33	0	0
15	1	8.16	12.05	33.3	0.5	0	0.5	0.5	1	1.5	1.5	0.5	1	2	0.5	0.5
15	2	4.5	8.16	100	0	0.33	0.33	0.33	0.67	1	1.33	1	0.67	2	0	0
15	3	8.7	4.5	50	0	0	0	0	0	0	0	1	0	0	0	0
15	4	6.03	8.7	33.3	0	0.5	0.5	0	0	0	2.5	0.5	0.5	1	0.5	0
15	5	9.7	6.03	33.3	0	0	0	0.67	0.67	1.33	1.67	0.33	0	1	0	0.33
16	1	12.3	11.5	33.3	0.33	0.67	1	0	1	1	1.33	1.67	0.33	1	0.33	0.33
16	2	9.55	12.3	50	0.67	0	0.67	1	1	2	1.67	1.67	0.67	2.33	0.67	0
16	3	11.53	9.55	66.7	0.5	0.5	1	0	2	2	1.5	1	0	3.5	0.5	0.5
16	4	9.5	11.53	100	0.33	0.33	0.67	0.67	1	1.67	2	1	0.67	3	0	0
16	5	11.8	9.5	50	0.5	0	0.5	0.5	1	1.5	2	1.5	1.5	3	1	1
17	1	4	9.46	100	0.33	0.33	0.67	0	0.67	0.67	1	0.33	0.33	1.33	0.33	0.33
17	2	11.6	4	66.7	0	0	0	0	0	0	2	0.5	0	2.5	0	0.5
17	3	4.87	11.6	33.3	0.67	0	0.67	0.67	1.33	2	0.33	0.67	0.33	2	0.67	0
17	4	6.4	4.87	0	0	0	0	0	0.33	0.33	1.33	0.33	0	3	0	0
17	5	9.86	6.4	66.7	0	0	0	0	0	0	2	1	1	2	0	0
18	1	10.03	4.1	33.3	0	0	0	0	0.5	0.5	2	3.5	1.5	1	0	0.5
18	2	8.35	10.03	0	0.33	0	0.33	0	0.67	0.67	3	1	1.33	1.67	0	0.33
18	3	14.5	8.35	66.7	0	0	0	1	2	3	1	0.5	0.5	1.5	0	0
18	4	18.8	14.5	66.7	0.67	0.33	1	1.33	0.67	2	4	2.33	1	1	0.67	0.67
18	5	8.9	18.8	50	0.67	0.33	1	0.67	2	2.67	4.67	1	2	2	0.33	0.33
19	1	6.96	6.85	33.3	0	0	0	0	2	2	0.5	1.5	0	2.5	0	0
19	2	8.63	6.96	66.7	0	0	0	0	1	1	1	1.33	1	1.67	0	0.33

Continua na próxima página

Tabela A.1: Continuação

id	rod	pont	pont_ult	local	gol	ass	part	fd	ff	ftotal	fs	fc	i	pe	rb	cart
19	3	11.25	8.63	50	0.33	0	0.33	0	0.33	0.33	0.67	0.67	0.67	1	0	0
19	4	12.75	11.25	50	0	0.5	0.5	1	1.5	2.5	1.5	0	0.5	0.5	0	0
19	5	11.9	12.75	33.3	0.5	0.5	1	0.5	0.5	1	2	0.5	1	2.5	0.5	0.5
20	1	7.56	5.86	33.3	0.33	0	0.33	0	0	0	0.67	2.33	0	1	0	0.33
20	2	9.25	7.56	50	0	0	0	0.67	1.33	2	1	1	0.67	1	0	0
20	3	2.94	9.25	66.7	0.5	0	0.5	0	0	0	1	1.5	1	0	0	0.5
20	4	11.93	2.94	66.7	0	0	0	0.33	0.33	0.67	1	3.67	0.33	1.33	0.33	0.67
20	5	5.46	11.93	33.3	0	1	1	0.67	0.67	1.33	2	0.67	0.33	1.33	0.33	0
21	1	7.3	7.6	33.3	0	0	0	0	2	2	1	2	1	0	0	0
21	2	6.1	7.3	33.3	0.33	0	0.33	0.67	0	0.67	0.67	1.33	0.33	1.67	0.33	0.33
21	3	3.2	6.1	66.7	0	0	0	0.33	0.33	0.67	0.67	1.33	1	1.33	0.67	0
21	4	7.43	3.2	33.3	0	0	0	0.67	0.33	1	1	1.67	0	0.67	0.67	0.33
21	5	4.74	7.43	66.7	0.33	0	0.33	0	0.67	0.67	1.33	2.67	0.33	0.33	1	0
22	1	4.5	5.7	100	0	0	0	0.5	0	0.5	1.5	0.5	0	0.5	0.5	0
22	2	7.36	4.5	33.3	0	0	0	1	0	1	2.5	2.5	0.5	1.5	0	1
22	3	5.8	7.36	100	0	0	0	0.33	1	1.33	3	0.33	0.33	1.67	0.33	0
22	4	9.83	5.8	66.7	0	0	0	0	0	0	4	3	1	4	0	0
22	5	6.23	9.83	33.3	0.67	0	0.67	0	0.67	0.67	1.33	0.33	0	0.67	0	0.67
23	1	10.16	3.55	66.7	0	0	0	0.5	0.5	1	1.5	1.5	0	1.5	0	1
23	2	5.9	10.16	66.7	0.33	0.33	0.67	1.33	0.33	1.67	2	2	0	0.67	0.33	0.33
23	3	3.3	5.9	50	0.33	0	0.33	0	0.67	0.67	1.33	3.33	0.33	2.67	0.67	0.33
23	4	5.5	3.3	100	0	0	0	0	1	1	1	3	0.5	2.5	1	0.5
23	5	5	5.5	33.3	0	0	0	0.5	1.5	2	1.5	1.5	0	1	0	0.5
24	1	11.73	7.36	66.7	0.33	0	0.33	0	0	0	0.33	0.33	0	1	0	0
24	2	5.16	11.73	66.7	0.33	0.33	0.67	0.33	0.67	1	2	0.67	1	2.33	0	0
24	3	11.4	5.16	66.7	0	0	0	0.67	0.33	1	0.67	1.67	0	1.33	0	0
24	4	9.76	11.4	33.3	0.33	0.33	0.67	0	0.67	0.67	0.67	0	1	1	0	0
24	5	11.63	9.76	66.7	0.33	0	0.33	0	0.67	0.67	2.33	0.67	1	0.33	0	0.33
25	1	8.2	7.86	100	0	0	0	1.33	1	2.33	1.67	0.33	0.33	2.33	0	0
25	2	8.9	8.2	50	0	0	0	2	0	2	5	0	0	4	1	0
25	3	13.06	8.9	33.3	0	0.5	0.5	0	0.5	0.5	1.5	1	0.5	0	0	0
25	4	9.2	13.06	66.7	0.67	0	0.67	0.67	0	0.67	3	0.33	0.67	1.33	0	0
25	5	6.63	9.2	33.3	0	0.33	0.33	0	0	0	4.33	0.33	0.67	1	0.33	0
26	1	7.8	6.6	100	0	0	0	1.5	0.5	2	2	0	0.5	4.5	0	0.5
26	2	7.3	7.8	50	0	0	0	1	0	1	0	1	2	3	0	0
26	3	6.4	7.3	0	0	0.5	0.5	0	0.5	0.5	0.5	0	0.5	2	0	0.5
26	4	4.7	6.4	0	0	0	0	1	0	1	1	0	0	1	0	0
26	5	5.93	4.7	66.7	0	0	0	1	0	1	1	0	0	0	0	1
27	1	5.35	11.7	100	0.5	0.5	1	0.5	0	0.5	0	1.5	1	0.5	0	0.5
27	2	4.5	5.35	66.7	0	0	0	0	0.5	0.5	1.5	0.5	0	1	0.5	0
27	3	10.2	4.5	66.7	0	0	0	0.33	0.33	0.67	0.67	1	0	3.33	0	0
27	4	4.94	10.2	33.3	0.33	0	0.33	0.33	0.33	0.67	1.33	1	1.67	2.67	0	0
27	5	3.45	4.94	50	0	0	0	0.33	0	0.33	0.33	1	1	2.67	0.33	0.33
28	1	8.26	11.25	33.3	0.5	0	0.5	0.5	1.5	2	0	1	1	1	0.5	0
28	2	10.85	8.26	50	0.33	0	0.33	0	0.67	0.67	1.67	2	0.67	2	0.33	0
28	3	8.16	10.85	66.7	0.5	0	0.5	0	1	1	0.5	0.5	1	1.5	0	0

Continua na próxima página

Tabela A.1: Continuação

id	rod	pont	pont_ult	local	gol	ass	part	fd	ff	ftotal	fs	fc	i	pe	rb	cart
28	4	8.55	8.16	50	0.33	0	0.33	0	0	0	1	0	0	0	0	0
28	5	6.05	8.55	100	0.5	0	0.5	0	0.5	0.5	1	1.5	1.5	2	0.5	0
29	1	6.56	5.3	66.7	0	0	0	0.33	0.67	1	2	3	1	1	1.33	0.33
29	2	5.05	6.56	100	0	0	0	1	0.67	1.67	0	1	0	1.67	0.33	0
29	3	6.96	5.05	33.3	0	0	0	0.5	1.5	2	0.5	2	0.5	2.5	0	0.5
29	4	6.76	6.96	66.7	0	0	0	0.67	0.67	1.33	1	1.67	1	0.67	0.67	0
29	5	4.85	6.76	0	0	0	0	0.33	0.33	0.67	1.67	1	1	1.33	0.67	0
30	1	8.23	6.8	33.3	0	0.33	0.33	0	0.67	0.67	0.67	1.33	0.67	3	0.33	0
30	2	7.35	8.23	50	0	0.33	0.33	0	1.33	1.33	0.33	1.67	2	1.67	1.33	0.33
30	3	4.35	7.35	50	0	0	0	1.5	1.5	3	1	1	0	0.5	1	0
30	4	5.7	4.35	50	0	0	0	0.5	0	0.5	0	0	0	0	0.5	0.5
30	5	5.13	5.7	33.3	0	0	0	0.5	0	0.5	1	0.5	0	0.5	0	0
31	1	9.85	9.76	50	0.33	0	0.33	0	1.67	1.67	2.67	1.33	1	1.33	0.67	0.33
31	2	3	9.85	100	0.5	0	0.5	0	1	1	0.5	1	0.5	1.5	0	0
31	3	6.25	3	100	0	0	0	2	0	2	0	3	0	3	0	1
31	4	6.4	6.25	0	0	0	0	1	0.5	1.5	2.5	1.5	0.5	2	0	0.5
31	5	5.96	6.4	33.3	0	0	0	1	1	2	2	2	0	2	0	0

Apêndice B

Códigos de programação

```
########################## BIBLIOTECAS ##########################

library(readxl)
library(nortest)
library(ggplot2)
library(geepack)
library(MASS)
library(corrplot)

########################## BANCO DE DADOS ##########################

banco <- read_excel("C:/Users/EDVALDO/Desktop/banco.xlsx")
View(banco)

########################## BANCO DE DADOS NOVO ##########################

banco.novo <- read_excel("C:/Users/EDVALDO/Desktop/banco.novo.xlsx")
View(banco.novo)

########################## ANÁLISE DESCRITIVA ##########################

 ### HISTOGRAMA DE PONT ###
```

```
ggplot(banco, aes(x = pont)) +
  geom_histogram(aes(y = ..count../sum(..count..), fill = ..count..),
                 binwidth = 1,color = 'white') +
  scale_fill_gradient("Freq. abs.", low = "gray85", high = "gray14") +
  scale_x_continuous(name = "Pontuação", breaks = seq(-5, 30, 5),
                     limits = c(-5,30)) +
  scale_y_continuous(name = "Frequência relativa", breaks = seq(0, 1, 0.025),
                     labels = scales::percent) +
  theme_minimal()

### MEDIDAS RESUMO DE PONT ###

summary(pont)
var(pont)

### ORIGEM DE PONT_MOD E PONT_ULT_MOD ###

pont_mod = pont + 5
pont_ult_mod = pont_ult + 5
banco = data.frame(banco, pont_ant, pont_ult_mod)

### HISTOGRAMA DE PONT_MOD ###

ggplot(banco, aes(x = pont_mod)) +
  geom_histogram(aes(y = ..count../sum(..count..), fill = ..count..),
                 binwidth = 1,color = 'white') +
  scale_fill_gradient("Freq. abs.", low = "gray85", high = "gray14") +
  scale_x_continuous(name = "Pontuação", breaks = seq(-5, 35, 5),
                     limits = c(-5,35)) +
  scale_y_continuous(name = "Frequência relativa", breaks = seq(0, 1, 0.025),
                     labels = scales::percent) +
  theme_minimal()
```

```
### CORRELAÇÕES E SEUS TESTES REFERENTES AS VARIÁVEIS
PONT_MOD E PONT_ULT_MOD ###

cor(pont_mod,pont_ult_mod, method = "pearson")
cor(pont_mod,pont_ult_mod, method = "spearman")
cor(pont_mod,pont_ult_mod, method = "kendall")

cor.test(pont_mod,pont_ult_mod, method = "pearson")
cor.test(pont_mod,pont_ult_mod, method = "spearman")
cor.test(pont_mod,pont_ult_mod, method = "kendall")

### MIGRAÇÃO PRO BANCO DE DADOS NOVO ###
### FACILITANDO A MANIPULAÇÃO DAS VARIÁVEIS###

pont = as.numeric(banco.novo$pont)
pont_ult = as.numeric(banco.novo$pont_ult)

### HISTOGRAMA DE PONT ###

ggplot(banco.novo, aes(x = pont)) +
  geom_histogram(aes(y = ..count../sum(..count..), fill = ..count..),
                 binwidth = 1,color = 'white') +
  scale_fill_gradient("Freq. abs.", low = "gray85", high = "gray14") +
  scale_x_continuous(name = "Pontuação média agrupando 3 rodadas",
                     breaks = seq(0, 25, 5), limits = c(0,25)) +
  scale_y_continuous(name = "Frequência relativa", breaks = seq(0, 1, 0.025),
                     labels = scales::percent) +
  theme_minimal()

### MEDIDAS RESUMO DE PONT ###

summary(pont)
var(pont)
```

```
### CORRELAÇÕES E SEUS TESTES REFERENTES AS VARIAVEIS PONT E PONT_ULT ###

cor(pont,pont_ult, method = "pearson")
cor(pont,pont_ult, method = "spearman")
cor(pont,pont_ult, method = "kendall")

cor.test(pont,pont_ult, method = "pearson")
cor.test(pont,pont_ult, method = "spearman")
cor.test(pont,pont_ult, method = "kendall")

### BANCO DE DADOS UTILIZADO PARA AS DESCRITIVAS DAS PREDITORAS E AJUSTE ###

banco_aj <- read.csv2("C:/Users/EDVALDO/Desktop/banco_aj.csv",
stringsAsFactors = F)

### TORNANDO AS VARIÁVEIS NUMÉRICAS ###

banco_aj$pont <- as.numeric(banco_aj$pont)
banco_aj$pont_ult <- as.numeric(banco_aj$pont_ult)
banco_aj$local <- as.numeric(banco_aj$local)
banco_aj$gol <- as.numeric(banco_aj$gol)

# o mesmo foi feito para as demais variáveis preditoras

### HISTOGRAMAS DAS VARIÁVEIS PREDITORAS ###

ggplot(banco_aj, aes(x = pont_ult)) +
  geom_histogram(aes(y = ..count../sum(..count..), fill = ..count..),
                 binwidth = 1,color = 'white') +
  scale_fill_gradient("Freq. abs.", low = "gray85", high = "gray14") +
  scale_x_continuous(name = "Pontuação média nas últimas 3 rodadas",
                     breaks = seq(0, 100, 5), limits = c(0,20)) +
```

```
  scale_y_continuous(name = "Frequência relativa", breaks = seq(0, 1, 0.025),
                     labels = scales::percent) +
  theme_minimal()

ggplot(banco_aj, aes(x = local)) +
  geom_histogram(aes(y = ..count../sum(..count..), fill = ..count..),
                 binwidth = 25, color = 'white') +
  scale_fill_gradient("Freq. abs.", low = "gray85", high = "gray14") +
  scale_x_continuous(name = "Porcentagem de jogos disputados dentro de casa
                     agrupando 3 rodadas", breaks = seq(0, 100, 25)) +
  scale_y_continuous(name = "Frequência relativa", breaks = seq(0, 1, 0.05),
                     labels = scales::percent) +
  theme_minimal()

ggplot(banco_aj, aes(x = gol)) +
  geom_histogram(aes(y = ..count../sum(..count..), fill = ..count..),
                 binwidth = 0.25,color = 'white') +
  scale_fill_gradient("Freq. abs.", low = "gray85", high = "gray14") +
  scale_x_continuous(name = "Quantidade média de gols agrupando 3 rodadas",
                     breaks = seq(0, 5, 0.25)) +
  scale_y_continuous(name = "Frequência relativa", breaks = seq(0, 1, 0.1),
                     labels = scales::percent) +
  theme_minimal()

# os histogramas das demais variáveis preditoras foram feitos da mesma maneira

### DIAGRAMAS DE DISPERSÃO DAS VARIÁVEIS PREDITORAS ###

ggplot(data = banco_aj, aes(x = pont_ult, y = pont)) +
  geom_point()+
  xlab("Pontuação média nas últimas 3 rodadas") +
  ylab("Pontuação média agrupando 3 rodadas") +
  theme_minimal()
```

```
ggplot(data = banco_aj, aes(x = local, y = pont)) +
  geom_point()+
  xlab("Porcentagem de jogos disputados dentro de casa agrupando 3 rodadas") +
  ylab("Pontuação média agrupando 3 rodadas") +
  theme_minimal()

ggplot(data = banco_aj, aes(x = gol, y = pont)) +
  geom_point()+
  xlab("Quantidade média de gols agrupando 3 rodadas") +
  ylab("Pontuação média agrupando 3 rodadas") +
  theme_minimal()

# os diagramas de dispersão das demais variáveis foram feitos da mesma maneira

### ACHANDO A MÉDIA DE PONT EM RELAÇÃO A CADA VALOR
QUE AS VARIÁVEIS PREDITORAS ASSUMEM ##

tabela1 = data.frame(banco_aj$local,banco_aj$pont)
banco.local = aggregate(tabela1[,2],list(banco_aj$local),mean)

tabela2 = data.frame(banco_aj$gol,banco_aj$pont)
banco.gol = aggregate(tabela2[,2],list(banco_aj$gol),mean)

tabela3 = data.frame(banco_aj$ass,banco_aj$pont)
banco.ass = aggregate(tabela3[,2],list(banco_aj$ass),mean)

# repetiu-se esse procedimento da mesma forma para as variáveis preditoras restantes

### DIAGRAMAS DE DISPERSÃO DAS VARIÁVEIS
PREDITORAS CONSIDERANDO A MÉDIA DE PONT ###

ggplot(data = banco.local, aes(x = Group.1, y = x)) +
```

```
  geom_point()+
  xlab("Porcentagem de jogos disputados dentro de casa agrupando 3 rodadas") +
  ylab("Média da variável pont") +
  theme_minimal()

ggplot(data = banco.gol, aes(x = Group.1, y = x)) +
  geom_point()+
  xlab("Quantidade média de gols agrupando 3 rodadas") +
  ylab("Média da variável pont") +
  theme_minimal()

ggplot(data = banco.ass, aes(x = Group.1, y = x)) +
  geom_point()+
  xlab("Quantidade média de assistências para gols agrupando 3 rodadas") +
  ylab("Média da variável pont") +
  theme_minimal()

# os demais diagramas considerando a média de pont foram feitos
da mesma maneira para as variáveis preditoras restantes

### MATRIZ DE CORRELAÇÃO ###

knitr::kable(cor(banco.novo))

### COR PLOT ###

C <- cor(banco_aj[,-c(1,2)]) # retirando as variáveis id e rod
corrplot(C, type = "upper")

############################## MODELAGEM ##############################

# os ajustes dos modelos via EEGs foram feitos através do pacote geepack,
conforme foi descrito na Subseção 4.3.1
```

```
### AJUSTES VIA MLGS ###
### EXEMPLO DISTRIBUIÇÃO NORMAL STEPWISE AIC ###

pont_normal = banco_aj$pont - 5
pont_ult_normal = banco_aj$pont_ult - 5

aj_normal = glm(pont_normal ~ pont_ult_normal + local + gol + ass + part +
ff + fd + ftotal + fs + fc + i + pe + rb + cart,
family = gaussian (link = "identity"), data = banco_aj)

# os ajustes são feitos dessa maneira para todos os modelos, variando as
variáveis presentes no modelo, a distribuição escolhida e a função de ligação

aj_normal_aic = stepAIC(aj_normal,direction = "both")

# para fazer o stepwise AIC através da função stepAIC mostrada acima
é preciso fazer o ajuste do modelo considerando todas as variáveis
preditoras existentes, conforme foi feito no aj_normal

### PODER PREDITIVO ###
# o banco de dados é carregado da mesma meneira feita anteriormente

### MODELO ALTERNATIVO ###

eqm_alt=mean((banco_pred$pont - banco_pred$pont_ult)^2)
eam_alt=mean(abs(banco_pred$pont - banco_pred$pont_ult))

# nos casos em que a f.l é a identidade é feito da mesma forma apresentada acima,
mas no lugar do termo "banco_pred$pont_ult" entra um vetor com os valores preditos
da variável resposta no ajuste em questão. Já nos casos em que a F.L é a logarítmica
é preciso passar a exponencial nos valores preditos com a função "exp()" antes
de colocar o vetor nas fórmulas das medidas (EQM e EAM) dos poderes preditivos
```