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

# A ESTATÍSTICA NO FUTEBOL: UMA ANÁLISE DOS PRINCIPAIS FATORES QUE INFLUENCIAM O NÚMERO DE GOLS FEITOS PELOS JOGADORES NO CAMPEONATO INGLÊS

**Robson Alves Mangerona**

**Trabalho de Conclusão de Curso**

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

# A estatística no futebol: uma análise dos principais fatores que influenciam o número de gols feitos pelos jogadores no campeonato inglês

**Robson Alves Mangerona**
**Orientadora: Profª.Drª. Daiane Aparecida Zuanetti**

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

**São Carlos**
**Fevereiro de 2023**

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

# Statistics in soccer: an analysis of the main factors that influence the number of goals scored by players in the Premier League

**Robson Alves Mangerona**
**Advisor: Profª.Drª. Daiane Aparecida Zuanetti**

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

**São Carlos**
**February 2023**

Robson Alves Mangerona

# A estatística no futebol: uma análise dos principais fatores que influenciam o número de gols feitos pelos jogadores no campeonato inglês

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Robson Alves Mangerona e aprovado pela banca examinadora.

Aprovado em 26 de Janeiro de 2023

Banca examinadora:

- Daiane Aparecida Zuanetti (Orientadora)
- Afrânio Márcio Corrêa Vieira
- Rafael Bassi Stern

*Dedico este trabalho a minha mãe, que está sempre ao meu lado em qualquer situação, a minha namorada pelo apoio incondicional, ao meu irmão e a todos que fizeram parte deste ciclo. E, principalmente, ao meu pai, que me orientou a ser a minha melhor versão.*

# Agradecimentos

Agradeço ao meu pai, Islanerson, que já se foi, mas que nunca deixou de ser a minha maior inspiração. À minha mãe, Alessandra, que é o meu maior exemplo de honestidade e bondade. À minha namorada, Amanda, que se tornou a maior parceira que eu poderia ter. Ao meu irmão, Roger, que desde que veio ao mundo é uma das minhas maiores motivações a ser uma pessoa cada vez melhor. E, por fim, aos demais familiares, professores e amigos que caminharam juntos comigo nessa longa e difícil estrada, mas que teve um final feliz.

# Resumo

O futebol é considerado um dos esportes mais populares do mundo e tem sua relevância refletida não apenas nas multidões que enchem os estádios ao redor dos 5 continentes dentro dos mais variados campeonatos, mas também se mostra de extrema importância na formação cultural e econômica de um país. Levando isso em consideração, com o passar do tempo, o futebol vem recebendo uma grande atenção de profissionais que tentam, por meio de análises estatísticas, auxiliar técnicos, dirigentes e, principalmente, jogadores para que estes encontrem, nos números, um caminho para alcançar suas melhores performances dentro de campo.

O objetivo desse trabalho é, portanto, analisar e identificar quais aspectos (número total de chutes, chutes a gol a cada 90 minutos, distância média do gol de todas as finalizações, gols previstos, entre outros) mais impactam na quantidade de gols feitos por jogador em uma temporada do campeonato inglês (calendário futebolístico comumente associado a um total de 38 jogos realizados entre 20 clubes). A partir disso, é entender os principais pontos que influenciam a performance final do jogador, uma vez que a quantidade de gols feitos pelos atletas reflete diretamente no desempenho de um time dentro do campeonato.

**Palavras-chave**: *número de gols, campeonato inglês, regressão binomial negativa.*

# Abstract

Soccer is considered one of the most popular sports in the world and its relevance is reflected not only in the crowds that fill the stadiums around the 5 continents within the most varied championships, but also proves to be extremely important in the cultural and economic development of a country. Taking this into account, over time, soccer has been a subject of interest for professionals who try, through statistical analysis, to help coaches, managers and, mainly, players find, in numbers, a way to reach their best performances on the field.

The objective of this work is, therefore, to analyze and identify which features (total number of shots, shots on goal every 90 minutes, average goal distance of all shots, predicted goals, among others) have the most impact on the number of goals scored by player in a season of Premier League (soccer calendar commonly associated with a total of 38 games played between 20 teams). From this, we want to understand the main factors that influence the final performance of the player, since the number of goals scored by the athletes directly reflects on the performance of a team within the championship.

**Key words**: *number of goals, Premier League, negative binomial regression.*

# Lista de Figuras


3.1 Boxplot da quantidade de gols marcados. . . . . 39
3.2 Diagrama de dispersão da idade por gols marcados. . . . . 40
3.3 Diagrama de dispersão de chutes por gols marcados. . . . . 41
3.4 Diagrama de dispersão de chutes com direção ao gol por gols marcados. . . 41
3.5 Diagrama de dispersão da porcentagem de chutes com direção ao gol por gols marcados. . . . . 42
3.6 Diagrama de dispersão de chutes em 90 minutos por gols marcados. . . . . 42
3.7 Diagrama de dispersão de chutes com direção ao gol em 90 minutos por gols marcados. . . . . 43
3.8 Diagrama de dispersão de gols por chutes por gols marcados. . . . . 44
3.9 Diagrama de dispersão de gols por chutes com direção ao gol por gols marcados. . . . . 44
3.10 Diagrama de dispersão de distância média do gol de todas as finalizações por gols marcados. . . . . 45
3.11 Diagrama de dispersão de chutes de falta por gols marcados. . . . . 45
3.12 Diagrama de dispersão de pênaltis convertidos por gols marcados. . . . . 46
3.13 Diagrama de dispersão de pênaltis batidos por gols marcados. . . . . 46
3.14 Diagrama de dispersão de gols previstos por gols marcados. . . . . 47
3.15 Função de Autocorrelação da série. . . . . 48
3.16 Matriz de correlação. . . . . 49
3.17 Valores da diagonal da matriz de projeção para modelo *Stepwise*. . . . . 54
3.18 Distância de cook para cada observação para modelo *Stepwise*. . . . . 55
3.19 Preditor linear versus resíduos e variável dependente ajustada para modelo *Stepwise*. . . . . 57
3.20 Envelope com 14,81% dos pontos para fora para modelo *Stepwise*. . . . . 58

3.21 Valores da diagonal da matriz de projeção e distância de cook para modelo lasso. . . . . 59
3.22 Preditor linear versus resíduos e variável dependente ajustada para modelo lasso. . . . . 59
3.23 Envelope com 19,26% dos pontos para fora para modelo lasso. . . . . 60
3.24 Preditor linear versus resíduos e variável dependente ajustada para modelo com distribuição binomial negativa considerando função de ligação raiz quadrada. . . . . 61
3.25 Envelope com 71,48% dos pontos para fora para modelo com distribuição binomial negativa considerando função de ligação raiz quadrada. . . . . 62
3.26 Preditor linear versus resíduos e variável dependente ajustada para modelo com distribuição Poisson considerando função de ligação logarítmica. . . . 62
3.27 Envelope com 94,81% dos pontos para fora para modelo com distribuição Poisson considerando função de ligação logarítmica. . . . . 63
3.28 Preditor linear versus resíduos e variável dependente ajustada para modelo com distribuição Poisson considerando função de ligação raiz quadrada. . . 63
3.29 Envelope com 0,74% dos pontos para fora para modelo com distribuição Poisson considerando função de ligação raiz quadrada. . . . . 64

# Lista de Tabelas


2.1 Quantidades $w_i$ e $f_i$ para ligação logarítmica. . . . . 27

3.1 Descritiva da quantidade de gols marcados pelos jogadores na liga inglesa. 39
3.2 ANODEV. . . . . 51
3.3 Teste de Wald. . . . . 51
3.4 Método *Stepwise*. . . . . 52
3.5 Método lasso. . . . . 52
3.6 ANODEV Modelo *Stepwise*. . . . . 53
3.7 ANODEV Modelo lasso. . . . . 53
3.8 Valores das variáveis dos possíveis pontos influentes. . . . . 56
3.9 $VIF$ de cada variável preditora para modelo *Stepwise*. . . . . 58
3.10 $VIF$ de cada variável preditora para modelo lasso. . . . . 60
3.11 Valores de AIC dos modelos analisados . . . . . 65
3.12 Estimativas dos coeficientes de regressão no modelo Poisson e binomial negativa considerando função de ligação raiz quadrada. . . . . 65
3.13 Estimativa dos coeficientes de regressão do modelo Poisson e binomial negativa considerando função de ligação raiz quadrada sem variável PT. . . . 66

# Sumário


**1 Introdução** **21**

**2 Regressão binomial negativa** **23**

2.1 Estimação por máxima verossimilhança . . . . . 24

2.2 Componentes do modelo . . . . . 27

2.3 Análise do desvio e seleção de modelos . . . . . 28

2.3.1 Função desvio . . . . . 28

2.3.2 Teste de Hipóteses . . . . . 30

2.3.3 Análise de desvio (ANODEV) . . . . . 32

2.3.4 Seleção de variáveis . . . . . 33

**3 Dados da premier league 2020-2021 e resultados** **37**

3.1 Análise Descritiva . . . . . 38

3.2 Análise de correlação entre os jogadores e associação entre as variáveis preditoras . . . . . 47

3.3 Resultados . . . . . 50

3.3.1 Modelagem . . . . . 50

3.3.2 Análise de diagnóstico - modelo *Stepwise* . . . . . 54

3.3.3 Análise de diagnóstico - modelo lasso . . . . . 59

3.3.4 Modelos alternativos . . . . . 61

**4 Conclusão e estudos futuros** **67**

**Referências Bibliográficas** **69**

**A Códigos** **71**

# Capítulo 1

# Introdução

Rasmus Ankersen, codiretor de futebol do Brentford, um clube da segunda divisão do campeonato inglês, deu, em uma entrevista durante a temporada 2019-2020, o seguinte depoimento: "sua posição na tabela não é a melhor métrica para avaliar o sucesso. Futebol é um esporte de baixa pontuação, então, eventos aleatórios têm um grande impacto. A bola pode desviar, o juiz pode errar. O melhor time ganha com menos frequência do que em outros esportes".

Pautado nestes fatos, Ankersen decidiu manter no cargo o técnico do seu time mesmo após a pressão sofrida por torcedores que queriam sua demissão, uma vez que a equipe estava nas últimas posições do campeonato. Ao invés de demitir o técnico, o codiretor do Brentford, juntamente com sua equipe de tecnologia, analisou estatísticas envolvendo o desempenho que os jogadores do time obtiveram até então no campeonato.

Assim sendo, utilizando um indicador chamado xG ou "*Expected Goals*" (quantidade de gols estimada por jogador levando em consideração a probabilidade de ocorrência desses gols em situações propícias a isso dentro das partidas), Ankersen percebeu que sua equipe criava boas chances com possibilidades reais de gols que, até aquele momento, não estavam se concretizando e, portanto, fatores específicos estavam prejudicando o time como, por exemplo, interferência do árbitro, distância média das finalizações, dentre outros. Por fim, o técnico não foi demitido e, aprimorando a equipe a partir dos indicadores analisados, em dez jogos dentro do campeonato, a equipe ganhou seis, empatou três e perdeu apenas um (LAW, 2020).

O caso exposto acima exemplifica como, na prática, uma análise bem estruturada dos dados referentes aos resultados obtidos até então pode ocasionar em uma grande mudança de mentalidade de todos os envolvidos e, a partir de treinamentos aprimorados e

de um estudo constante, melhores resultados serem obtidos. Isto porque, como mostrado, quando o objetivo é saber o quão bem uma equipe provavelmente jogará no futuro, é mais recomendado olhar para as estatísticas subjacentes do que para sua posição na classificação (Caley, 2014).

Pensando nisso, este trabalho tem como objetivo fazer uma análise estatística em relação ao número de gols marcados por jogador com dados da temporada 2020-2021 do campeonato inglês. Dessa forma, será levado em consideração diversas variáveis que podem ou não influenciar nesse resultado, dentre eles o xG de cada jogador descrito anteriormente (StatsBomb, 2022).

Essa análise será realizada com o auxílio de métodos estatísticos mais sofisticados do que aqueles que normalmente são utilizados como, por exemplo, a regressão linear múltipla (da Silva, 2018). Dessa forma, a regressão binomial negativa (Cordeiro e Demétrio, 2008), baseada na distribuição binomial negativa, será explorada de modo a incorporar de maneira mais eficaz as necessidades dos dados que serão utilizados e do próprio objetivo proposto.

O relatório está organizado da seguinte maneira: o Capítulo 2 apresenta os procedimentos metodológicos que serão aplicados ao longo do trabalho para a obtenção dos objetivos definidos. Já o Capítulo 3 mostra como o banco de dados de interesse está estruturado, apresenta uma análise descritiva desse conjunto de dados e também os principais resultados de modelagem. Por fim, o Capítulo 4 contém a conclusão do trabalho e sugestão de estudos futuros.

# Capítulo 2

# Regressão binomial negativa

Como dito anteriormente, neste trabalho utilizaremos a metodologia da regressão binomial negativa para predizer o número de gols que um jogador do campeonato inglês faz e também identificar quais variáveis preditoras mais impactam nesse número. Nesse capítulo apresentaremos essa metodologia.

A regressão binomial negativa é utilizada para modelar variáveis de contagem que seguem uma distribuição binomial negativa e que apresentam uma alta variabilidade, situação na qual a regressão Poisson não é recomendável. Como a variável resposta do trabalho diz respeito a quantidade de gols que um jogador faz ao longo da temporada, tal evento pode variar muito de um jogador para o outro e, portanto, espera-se que a regressão binomial negativa apresente um comportamento mais adequado do que a regressão Poisson. Dessa forma, os possíveis valores da variável resposta $Y$ são os não negativos inteiros: 0, 1, 2, 3, ...

A regressão binomial negativa é uma generalização da regressão de Poisson que flexibiliza a suposição que a variância da variável resposta é igual a média. O modelo de regressão binomial negativa é baseado na distribuição da mistura Poisson-gama que assume para o parâmetro da distribuição Poisson uma distribuição gama (com dois parâmetros) e integra a distribuição obtida em relação a esse parâmetro (Cameron e Trivedi, 2013).

A função de probabilidade da distribuição da mistura Poisson-gama (binomial negativa) resultante é igual a

$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\phi}{\phi+\mu}\right)^{\phi}\left(\frac{\mu}{\phi+\mu}\right)^{y} \tag{2.1}$$

em que $y = 0, 1, 2, ...,\ \mu > 0$ e $\phi > 0$.

Portanto, $Y$ segue distribuição binomial negativa de média $\mu$ e parâmetro de dispersão $\phi > 0$ $(BN(\mu, \phi))$.

Suponha agora que $Y_1, ..., Y_n$ são variáveis aleatórias independentes tal que $Y_i$ tem distribuição $BN(\mu_i, \phi)$. A função de probabilidade de $Y_i$ é dada por:

$$f(y_i; \mu_i, \phi) = \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \mu_i}\right)^{\phi} \left(\frac{\mu_i}{\phi + \mu_i}\right)^{y_i}. \tag{2.2}$$

Temos que $\mathrm{E}(Y_i) = \mu_i$ e $\mathrm{Var}(Y_i) = \mu_i + \frac{(\mu_i)^2}{\phi}$. Como a distribuição binomial negativa quando o parâmetro $\phi$ é conhecido pertence à família exponencial e é um caso particular dos modelos lineares generalizados (Cordeiro e Demétrio, 2008), a parte sistemática é dada por $g(\mu_i) = \eta_i = x_i^T \beta$, em que $x_i = (x_{i1}, ..., x_{ip})^T$ contém valores de variáveis preditoras, $\beta = (\beta_1, ..., \beta_p)^T$ é um vetor de parâmetros desconhecidos, os coeficientes de regressão, e $g(.)$ é a função de ligação. Mais detalhes são mostrados a seguir.

## 2.1 Estimação por máxima verossimilhança

Os coeficientes de regressão apresentados anteriormente podem ser estimados através do método de máxima verossimilhança e, pensando nisso, podemos definir $\theta = (\beta^T, \phi)^T$. Para estimar $\theta$ pelo método da máxima verossimilhança precisamos, primeiramente, definir a função de verossimilhança, que nesse caso é dada por

$$L(\theta) = \prod_{i=1}^{n} \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \mu_i}\right)^{\phi} \left(\frac{\mu_i}{\phi + \mu_i}\right)^{y_i}. \tag{2.3}$$

Aplicando o logaritmo na função anterior, chegamos em:

$$l(\theta) = \sum_{i=1}^{n} \left[ \log \left\{ \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \right\} + \phi \log \phi + y_i \log \mu_i - (\phi + y_i) \log(\mu_i + \phi) \right] \tag{2.4}$$

em que $\mu_i = g^{-1}(x_i^T \beta)$.

O estimador de máxima verossimilhança para $\theta$ é o valor que maximiza a Equação (2.4) e para determiná-lo precisamos da função escore. A fim de obter a função escore

para $\beta$, calcula-se as seguintes derivadas:

$$\begin{aligned}\frac{\partial l(\theta)}{\partial \beta_j} &= \sum_{i=1}^{n}\left(\frac{y_i}{\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} - \frac{(\phi+y_i)}{(\phi+\mu_i)}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}\right)\\ &= \sum_{i=1}^{n}\left(\frac{y_i}{\mu_i}\frac{d\mu_i}{d\eta_i}x_{ij} - \frac{(\phi+y_i)}{(\phi+\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ij}\right)\\ &= \sum_{i=1}^{n}\left(\frac{\phi(d\mu_i/d\eta_i)}{\mu_i(\phi+\mu_i)}(y_i-\mu_i)x_{ij}\right)\\ &= \sum_{i=1}^{n} w_i f_i^{-1}(y_i-\mu_i)x_{ij},\end{aligned}$$

em que $w_i = (d\mu_i/d\eta_i)^2/(\mu_i^2\phi^{-1}+\mu_i)$ e $f_i = d\mu_i/d\eta_i$. Logo, é possível expressar a função escore na forma matricial a seguir:

$$U_\beta(\theta) = X^T W F^{-1}(y-\mu), \tag{2.5}$$

em que $X$ é a matriz modelo com linhas $x_i^T$, $i = 1, ..., n$, $W = diag\{w_1, ..., w_n\}$, $F = diag\{f_1, ..., f_n\}$, $y = (y_1, ..., y_n)^T$ e $\mu = (\mu_1, ..., \mu_n)^T$.

A função escore para $\phi$ é obtida de forma semelhante ao caso anterior e é dada por

$$U_\phi(\theta) = \sum_{i=1}^{n}[\psi(\phi+y_i) - \psi(\phi) - (y_i+\phi)/(\phi+\mu_i) + \log\{\phi/(\phi+\mu_i)\} + 1], \tag{2.6}$$

em que $\psi(.)$ é a função digama (derivada logarítmica da função gama).

Para obtermos a matriz de informação de Fisher calculamos as derivadas abaixo:

$$\frac{\partial^2 l(\theta)}{\partial\beta_j\partial\beta_l} = -\sum_{i=1}^{n}\left\{\frac{(\phi+y_i)}{(\phi+y_i)^2} - \frac{y_i}{\mu_i^2}\right\}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij}x_{il} + \sum_{i=1}^{n}\left\{\frac{y_i}{\mu_i} - \frac{(\phi+y_i)}{(\phi+\mu_i)}\right\}\frac{d^2\mu_i}{d\eta_i^2}x_{ij}x_{il}, \tag{2.7}$$

cujos valores esperados são dados por

$$\begin{aligned}E\left\{\frac{\partial^2 l(\theta)}{\partial\beta_j\partial\beta_l}\right\} &= -\sum_{i=1}^{n}\frac{\phi(d\mu_i/d\eta_i)^2}{(\phi+\mu_i)}x_{ij}x_{il}\\ &= -\sum_{i=1}^{n} w_i x_{ij}x_{il}.\end{aligned}$$

Dessa forma, podemos expressar a informação de Fisher para $\beta$ da seguinte forma

matricial:

$$K_{\beta\beta}(\theta) = E\left\{-\frac{\partial^2 l(\theta)}{\partial\beta\partial\beta^T}\right\} = X^T W X. \tag{2.8}$$

Ademais, segundo Lawless (1987), a informação de Fisher para $\phi$ é dada por:

$$K_{\phi\phi}(\theta) = \sum_{i=1}^{n}\left\{\sum_{j=0}^{\infty}(\phi+j)^{-2}Pr(Y_i \geq j) - \phi^{-1}\mu_i/(\mu_i+\phi)\right\}. \tag{2.9}$$

Lawless (1987) diz também que $\beta$ e $\phi$ são parâmetros ortogonais e, portanto, a matriz de informação de Fisher para $\theta$ assume a forma bloco diagonal:

$$K_{\theta\theta} = \begin{bmatrix} K_{\beta\beta} & 0 \\ 0 & K_{\phi\phi} \end{bmatrix}. \tag{2.10}$$

As estimativas de máxima verossimilhança para $\beta$ e $\phi$ podem ser obtidas através de um algoritmo de mínimos quadrados reponderados, aplicando o método escore de Fisher, a partir da Equação (2.5) e do método de Newton-Raphson para obter $\hat{\phi}$ desenvolvido a partir da Equação (2.6), os quais são descritos como

$$\beta^{m+1} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} y^{*(m)} \tag{2.11}$$

e

$$\phi^{m+1} = \phi^{(m)} - \left\{U_{\phi}^{(m)}/\ddot{L}_{\phi\phi}^{(m)}\right\}, \tag{2.12}$$

para $m = 0, 1, 2, ...$ em que

$$y^{*(m)} = X\beta^{(m)} + F^{-1}(y - \mu^{(m)}) \tag{2.13}$$

é uma variável dependente modificada e a derivada da Equação (2.6) em relação a $\phi$ é dada por:

$$\ddot{L}_{\phi\phi} = \sum_{i=1}^{n}\left\{\frac{d\psi(\phi+y_i)}{d\phi} + (y_i - 2\mu_i - \phi)/(\phi+\mu_i)^2\right\} + n\phi^{-1}\left\{1 - \phi\frac{d\psi(\phi)}{d(\phi)}\right\}. \tag{2.14}$$

Os dois procedimentos apresentados acima são aplicados simultaneamente até a convergência (Paula, 2004).

## 2.2 Componentes do modelo

Como abordado anteriormente, os modelos lineares generalizados (MLG) apresentam, de uma forma geral, três principais componentes que são:

- **Componente aleatório:** conjunto de variáveis aleatórias independentes $Y_1, ..., Y_n$ provenientes de uma mesma distribuição com médias $\mu_1, ..., \mu_n$;
- **Componente sistemático:** variáveis explicativas que entram no modelo na forma de uma soma linear de seus efeitos; e
- **Função de ligação:** uma função contínua, diferenciável e monótona que relaciona o componente aleatório ao componente sistemático, ou seja, vincula a média ao preditor linear.

No caso do presente trabalho, sabemos que o componente aleatório do modelo, a variável resposta, é o número de gols feitos por cada jogador na temporada e, portanto, se trata de uma distribuição binomial negativa para dados de contagem.

Além disso, como componente sistemático, temos $x_i^T\beta$ com $x_i^T$ sendo os valores das variáveis preditoras presentes no conjunto de dados de interesse para o jogador $i$ da liga inglesa 2020-2021 e $\beta$ os coeficientes de regressão que queremos estimar.

Por fim, precisamos ter uma função de ligação adequada que relacione os dois componentes acima e essa ligação tem objetivos importantes diante do modelo como o de linearizar a relação entre $\mu_i$ e $\eta_i$, produzir valores válidos (pertencentes ao espaço paramétrico) de $\mu_i$ para qualquer conjunto de valores para as variáveis preditoras e proporcionar interpretações práticas para os coeficientes (parâmetros) presentes no preditor linear.

Para o caso de $Y$ com distribuição binomial negativa, a função mais comumente utilizada e que também será acionada para este trabalho é a ligação logarítmica. A Tabela 2.1 abaixo apresenta as expressões para $w_i$ e $f_i$ para essa ligação usual em modelos com resposta binomial negativa (Cordeiro e Demétrio, 2008).

Tabela 2.1: Quantidades $w_i$ e $f_i$ para ligação logarítmica.

| **Ligação** | $w_i$ | $f_i$ |
|---|---|---|
| $\log(\mu_i) = \eta_i$ | $\mu_i/(\mu_i\phi^{-1} + 1)$ | $\mu_i$ |

Um recurso disponível para a estimação de um modelo de regressão binomial negativa e que será utilizado no trabalho é o pacote `MASS` do R que possibilita a obtenção das

estimativas dos coeficientes de regressão pelo método de máxima verossimilhança com o uso da função de ligação desejada através da função `glm.nb()`. Além desse pacote, utilizaremos o `car`. Ambos pacotes conduzem os testes de hipóteses, que serão abordados a seguir, associados a modelagem de interesse.

## 2.3 Análise do desvio e seleção de modelos

Levando em consideração que a variável resposta $Y$, quantidade de gols marcados por jogador na temporada, e a função de ligação logarítmica escolhida sejam uma boa combinação, o objetivo é definir quantos termos são necessários na estrutura linear para uma descrição adequada dos dados.

O ideal é encontrar um modelo que não seja nem muito complexo com uma quantidade muito grande de variáveis e, por outro lado, nem muito simples com uma quantidade muito pequena de variáveis. Dessa forma, um modelo intermediário com fácil interpretação e com um número adequado de variáveis que expliquem bem os dados é o mais desejado.

Nesse sentido, alguns métodos e técnicas devem ser utilizados para fazer a melhor escolha das variáveis que são relevantes ao modelo estudado. Pensando nisso, vamos introduzir alguns conceitos a seguir começando pela função desvio.

### 2.3.1 Função desvio

Considerando um conjunto de dados com $n$ observações, podemos ajustar modelos com até $n$ parâmetros. Então, temos algumas opções como o modelo nulo ($\beta_0$ é o único parâmetro) e o modelo saturado (possui $n$ parâmetros, um para cada observação). Enquanto que o modelo nulo é muito simples para explicar o conjunto de informações, o modelo saturado, na prática, é inapropriado, uma vez que não sumariza os dados, mas, simplesmente, os repete.

Além destes, temos também outros modelos que não são extremos como os anteriores, como o minimal (possui o menor número de termos necessário ao ajuste) e o maximal (possui o maior número de termos necessário ao ajuste). De uma forma geral, o modelo ideal é obtido com a inclusão sucessiva de termos ao modelo minimal até chegar ao maximal, o que é chamado de modelo encaixado. E, qualquer modelo com $p$ parâmetros linearmente independentes, situado entre os modelos minimal e maximal, é conhecido como modelo corrente (Cordeiro e Demétrio, 2008).

Dessa forma, o desafio é determinar a necessidade do incremento de um parâmetro no modelo corrente ou, então, verificar a falta de ajuste com sua omissão. Surge, portanto, o interesse na utilização de medidas de discrepância que medem o ajuste de um modelo como o desvio, ou *deviance* do inglês, que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado e do modelo corrente avaliado na estimativa de máxima verossimilhança $\hat{\beta}$.

Seja o logaritmo da função de verossimilhança definido por:

$$l(\mu; y) = \sum_{i=1}^{n} l(\mu_i; y_i), \tag{2.15}$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = x_i^T\beta$. Para o modelo saturado ($p = n$), a função $l(\mu; y)$ é definida por:

$$l(y; y) = \sum_{i=1}^{n} l(y_i; y_i). \tag{2.16}$$

Então, nesse caso, a estimativa de máxima verossimilhança (MV) de $\mu_i$ é $\tilde{\mu_i} = y_i$ e, quando $p < n$, a estimativa de $l(\mu; y)$ é dada por $l(\hat{\mu}; y)$ com a estimativa de MV de $\mu_i$ sendo $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ e $\hat{\eta}_i = x_i^T\hat{\beta}$. Dessa forma, a qualidade do ajuste de um modelo é avaliada pela função desvio definida como

$$D(y; \hat{\mu}) = 2\left\{l(y; y) - l(\hat{\mu}; y)\right\}. \tag{2.17}$$

Um valor pequeno para a função desvio indica que, para um modelo com menor número de parâmetros, o ajuste é tão bom quanto o ajuste com o modelo saturado.

Por fim, a função desvio no caso binomial negativo, assumindo $\phi$ fixo, é dada por:

$$D(y; \hat{\mu}) = 2\sum_{i=1}^{n}\left[\phi \log\left\{\frac{\hat{\mu}_i + \phi}{y_i + \phi}\right\} + y_i \log\left\{\frac{y_i(\hat{\mu}_i + \phi)}{\hat{\mu}_i(y_i + \phi)}\right\}\right], \tag{2.18}$$

em que $\hat{\mu}_i = g^{-1}(x_i^T\hat{\beta})$ (Paula, 2004).

Um resultado assintótico importante segundo Jørgensen (1987) é o de que, em geral, para os casos em que $D(y; \hat{\mu})$ depende do parâmetro de dispersão $\phi^{-1}$ temos que $D(y; \hat{\mu}) \sim \chi^2_{n-p}$, quando $\phi \to \infty$.

### 2.3.2 Teste de Hipóteses

Os testes de hipóteses associados ao vetor de parâmetros $\beta$ podem ser formulados em termos de hipóteses lineares na seguinte forma:

$$H_0 : C\beta = \xi \text{ versus } H_1 : C\beta \neq \xi.$$

Dado $M_p$ um modelo com $p$ parâmetros e $M_q$ um modelo encaixado a $M_p$, temos acima que $C$ é uma matriz $q$ x $p$, com $q \leq p$ e $\xi$ é um vetor de dimensão $q$ previamente especificado.

Geralmente, temos o interesse em testar dois tipos de hipóteses, são elas:

- Hipótese de nulidade de um componente do vetor de parâmetros que é dado por

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0,$$

para algum $j$, sendo neste caso $q = 1$, $C = (0, ..., 0, 1, 0, ..., 0)$ e ocupando o 1 a $j$-ésima posição e $\xi = 0$; ou

- Hipótese de nulidade de um subvetor do vetor de parâmetros que é dado por

$$H_0 : \beta_r = 0 \text{ versus } H_1 : \beta_r \neq 0,$$

para algum subvetor de $r$ componentes de $\beta$.

Será abordado a seguir os dois casos mais usados para testar a hipótese $H_0$. São os testes de razão de verossimilhança e o de Wald que serão aplicados ao conjunto de dados do trabalho (Turkman e Silva, 2000).

#### Teste de razão de verossimilhança (TRV)

A estatística de Wilks ou estatística de razão de verosimilhanças (TRV) é definida como

$$\begin{aligned} \Lambda &= -2 \log \frac{max_{H_0} L(\beta)}{max_{H_0 \cup H_1} L(\beta)} \\ &= -2 \left\{ l(\tilde{\beta}) - l(\hat{\beta}) \right\}, \end{aligned}$$

em que $\tilde{\beta}$ é o valor de $\beta$ que maximiza a verossimilhança sujeita às restrições da hipótese $C\beta = \xi$.

O teorema de Wilks estabelece que a estatística $\Lambda$ tem, sob $H_0$, uma distribuição assintótica $\chi^2$ sendo o número de graus de liberdade igual à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso $p$) e o número de parâmetros a estimar sob $H_0$ (neste caso $p - q$). Assim, sob $H_0$,

$$-2\left\{l(\tilde{\beta}) - l(\hat{\beta})\right\} \sim \chi^2_q.$$

A partir desse teste, a hipótese nula é rejeitada a favor da hipótese alternativa, a um nível de significância $\alpha$, se o valor observado da estatística $\Lambda$ for superior ao quantil de probabilidade $1 - \alpha$ de uma distribuição $\chi^2_q$.

**Teste de Wald**

Seja $\hat{\beta}$ o estimador de máxima verossimilhança de $\beta$, que possui uma distribuição assintótica $N_p(\beta, \mathcal{I}^{-\infty}(\hat{\beta}))$, com $\mathcal{I}^{-\infty}(\hat{\beta})$ sendo a inversa da matriz informação de Fisher com a substituição do vetor $\beta$ pela estimativa e admitindo que para grandes amostras $\mathcal{I}(\beta) \approx \mathcal{I}(\hat{\beta})$. Dado que o vetor $C\hat{\beta}$ é uma transformação linear de $\hat{\beta}$ então, pelas propriedades da distribuição normal multivariada,

$$C\hat{\beta} \sim N_q(C\beta, C\mathcal{I}^{-\infty}(\hat{\beta})\mathcal{C}^{\mathcal{T}}) \tag{2.19}$$

e, consequentemente, sob a hipótese nula, a estatística de Wald é dada por:

$$W = (C\hat{\beta} - \xi)^T [C\mathcal{I}^{-\infty}(\hat{\beta})\mathcal{C}^{\mathcal{T}}]^{-\infty}(\mathcal{C}\hat{\beta} - \xi), \tag{2.20}$$

que também possui uma distribuição assintótica $\chi^2$ com $q$ graus de liberdade.

Por fim, a hipótese nula é rejeitada, a um nível de significância $\alpha$, se o valor observado da estatística de Wald for superior ao quantil de probabilidade $1 - \alpha$ de uma distribuição $\chi^2$ (Turkman e Silva, 2000).

O teste de Wald é um teste simples e fácil de calcular com base apenas em estimativas de parâmetros e seus erros padrão (assintóticos). O teste de razão de verossimilhança, por outro lado, requer as verossimilhanças do modelo completo e do modelo reduzido. É computacionalmente mais exigente, mas também fornece o teste assintoticamente mais poderoso e confiável. O TRV é quase sempre preferível ao teste de Wald, a menos que demandas computacionais tornem impraticável reajustar o modelo (Fox, 1997).

### 2.3.3 Análise de desvio (ANODEV)

A análise de desvio ou simplesmente ANODEV é um caso generalizado da análise de variância para os MLG. Seu objetivo é, por meio de uma sequência de modelos encaixados, obter cada um desses modelos incluindo mais termos do que temos nos modelos anteriores, as variáveis preditoras e suas interações. O desvio, visto anteriormente, é usado como uma medida de discrepância do modelo e, dessa forma, é construída uma tabela de diferenças de desvios.

A tabela ANODEV é a representação de uma sequência de testes de razão de verossimilhanças para um modelo linear generalizado, em que os termos do preditor linear são acrescentados sucessivamente ao modelo (começando pelo modelo nulo), e a significância das inclusões é avaliada via TRV. A ordem de inclusão das variáveis no modelo é escolhida pelo usuário e, geralmente, a ordenação das variáveis acaba por alterar a significância das mesmas.

Sejam os modelos encaixados $M_q$ e $M_p$ ($M_q \subset M_p$, $q < p$), com $q$ e $p$ parâmetros, respectivamente. A estatística $D_q - D_p$ com $(p - q)$ graus de liberdade é interpretada como uma medida do quanto a variação dos dados é explicada pelos termos que estão em $M_p$ e não estão em $M_q$, incluídos os efeitos dos termos em $M_q$ e ignorando quaisquer efeitos dos termos que não estão em $M_p$. Usando as funções desvios, a estatística TRV, vista anteriormente, se escreve, para $\phi$ conhecido e usando propriedades assintóticas, como

$$\phi^{-1}(D_q - D_p) \sim \chi^2_{p-q}. \tag{2.21}$$

Se $\phi$ é desconhecido, deve-se obter uma estimativa $\hat{\phi}$ consistente baseada no modelo maximal com $w$ parâmetros necessários ao ajuste e a inferência pode ser baseada na estatística $F$, dada por (Cordeiro e Demétrio, 2008):

$$F = \frac{(D_q - D_p)/(p - q)}{\hat{\phi}} \sim F_{p-q,n-w}. \tag{2.22}$$

A fim de exemplificar a utilização da tabela ANODEV, considere o modelo linear generalizado abordado neste trabalho, com $Y$ tendo distribuição binomial negativa, e com 3 variáveis (Idade, TC e CaG) no preditor linear e incluídas no modelo nessa respectiva ordem. Então, a tabela ANODEV irá mostrar os desvios, as diferenças de desvios, os respectivos graus de liberdade e os TRVs associados para:

- Inclusão da variável Idade no modelo que possui apenas o intercepto;
- Inclusão da variável TC no modelo que já possui Idade; e
- Inclusão da variável CaG no modelo que já possui tanto Idade quanto TC.

Outra forma de construir a análise do desvio é avaliando a significância de uma variável quando incluída no modelo que contém todas as demais variáveis. No exemplo acima, a análise seria feita da seguinte maneira:

- Inclusão da variável Idade no modelo que já possui TC e CaG;
- Inclusão da variável TC no modelo que já possui Idade e CaG; e
- Inclusão da variável CaG no modelo que já possui Idade e TC.

### 2.3.4 Seleção de variáveis

Como vimos anteriormente, existem algumas alternativas para, dentre todos os possíveis modelos, encontrarmos um que melhor explica a variável resposta e os dados de uma forma geral e, portanto, nos fornece uma melhor predição do problema de interesse. Contudo, principalmente em casos nos quais o número de variáveis é muito grande, há procedimentos automáticos que selecionam, sequencialmente, um subconjunto de variáveis a serem incluídas no modelo de modo que, ao final do processo, as variáveis ainda presentes são aquelas que realmente são relevantes para explicar o problema analisado levando em consideração determinado critério.

Estes procedimentos foram desenvolvidos com o objetivo de economizar esforço computacional, uma vez que os métodos para encontrar o melhor modelo dentre todos os possíveis podem exigir muito esforço dependendo da complexidade desse modelo. Dessa forma, temos diversos métodos de seleção automática de variáveis e, as que serão abordadas neste projeto, são os métodos *Stepwise* pelo critério AIC e o lasso (Tibshirani, 1996), do inglês *least absolute shrinkage and selection operator*, para o caso binomial negativa (Neter *et al.*, 1989).

#### Stepwise

O procedimento conhecido por *Stepwise* é, provavelmente, um dos métodos mais utilizados com o propósito de seleção de variáveis. De uma forma geral, este método ajusta

uma sequência de modelos em que, em cada passo, adiciona ou deleta uma variável candidata do modelo corrente.

Outro método que também é muito utilizado é o *forward* (passo à frente) onde é ajustada uma sequência de modelos que em cada passo uma variável é adicionada. Temos também o *backward* (passo atrás) que, diferente do primeiro caso, é ajustada uma sequência de modelos que em cada passo uma variável é excluída. Por fim, temos o *Stepwise* passo a passo que ajusta uma sequência de modelos e em cada etapa uma variável é excluída ou adicionada, isto é, se trata de um processo alternado em relação à *forward* e *backward* onde é adicionada a variável mais significativa ou removida a variável menos significativa durante cada passo.

O critério para adicionar ou excluir uma variável pode ser definido através de várias métricas, como o AIC, por exemplo. O critério de AIC, proposto por Akaike (1974), se trata de um processo de minimização que não envolve testes estatísticos. O objetivo é selecionar um modelo que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança $l(\beta)$ cresce com o aumento do número de parâmetros do modelo, a proposta é encontrar o modelo com menor valor para a seguinte função:

$$AIC = -2l(\hat{\beta}) + 2p,$$

em que $p$ é o número de parâmetros. Para o caso onde o modelo tem resposta binomial negativa, o interesse é encontrar um submodelo para o qual a quantidade abaixo seja minimizada:

$$AIC = D(y; \hat{\mu}) + 2p, \tag{2.23}$$

em que $D(y; \hat{\mu})$ é a função desvio do modelo. Nesse caso, uma variável preditora é incluída no modelo se ela melhorar o valor de AIC do modelo corrente e uma variável é excluída no *backward* se o AIC do modelo melhorar com essa exclusão (Paula, 2004).

A função no R que deve ser acionada para conduzir o método acima é "stepAIC" (Venables e Ripley, 2013).

### Lasso

Um método alternativo ao *Stepwise* para seleção de variáveis é o lasso. Esse método, na maioria dos casos, possui um custo computacional menor em relação ao *Stepwise*, o que o torna uma técnica interessante.

A seleção de variáveis é realizada através do método de encolhimento, que gera estimativas de coeficientes que são muito próximas de zero. Além disso, o lasso reduz o erro da predição comparado ao estimador de máxima verossimilhança mas, por outro lado, é um estimador viciado para $\beta$.

Essa técnica inclui um termo de penalidade que restringe o tamanho dos coeficientes estimados. Nesse sentido, à medida que o termo de penalidade aumenta, o lasso define mais coeficientes iguais a zero. Isso significa que o estimador lasso é, em geral, um modelo menor, com menos variáveis.

Formalmente, no lasso, buscamos por:

$$\hat{\beta}_{L,\lambda} = \min_{\beta}(\frac{1}{n}D(y;\hat{\mu}) + \lambda\sum_{j=1}^{p}|\beta_j|), \tag{2.24}$$

em que $D(y;\hat{\mu})$ é a função desvio do modelo, $\sum_{j=1}^{p}|\beta_j|$ é o termo de penalidade (se $\beta_1$ corresponde ao intercepto do modelo, ele não deve ser considerado e o somatório começa em $j = 2$) e $\lambda$ é um parâmetro não negativo de regularização, geralmente escolhido através do método de validação cruzada, $k$-fold. Essa validação cruzada é um método no qual os dados são divididos em $k$ partes (subconjuntos) aproximadamente de mesmo tamanho. O modelo é estimado com $k-1$ partes e testado na única parte que não é utilizada para estimar o modelo. Para cada valor de $\lambda$, são propostos $k$ modelos e, para cada um, é calculado o erro de predição ($EP$) como

$$EP = \sum_{i}(y_i - \hat{y}_i)^2, \tag{2.25}$$

em que $y_i$ e $\hat{y}_i$ são, respectivamente, o valor observado e o valor predito das observações que pertencem ao subconjunto não utilizado no ajuste do modelo.

Esse procedimento é repetido $k$ vezes, alterando, a cada vez, os $k-1$ subconjuntos sobre os quais o modelo é estimado e, consequentemente, o subconjunto na qual ele é testado. Portanto, para cada valor de $\lambda$, temos $k$ erros de predição e calculamos a média desses $EPs$. Por fim, o valor de $k$ utilizado para análise é o valor padrão obtido da função no R e o $\lambda$ escolhido é aquele que apresenta a menor média de $EP$. À medida que $\lambda$ aumenta, o número de componentes diferentes de zero de $\beta$ diminui. (Tibshirani, 1996).

A função no R que deve ser acionada para conduzir o método lasso para o modelo binomial-negativa é `glmnet`. Para conduzir a validação cruzada para escolha do melhor

valor de $\lambda$ é a `cv.glmnet` (Friedman *et al.*, 2010).

Por fim, a proposta deste trabalho é aplicar os dois métodos de seleção de variáveis apresentados acima nos dados que estão sendo estudados e comparar os resultados que serão obtidos.

# Capítulo 3

# Dados da premier league 2020-2021 e resultados

O banco de dados utilizado neste trabalho foi disponibilizado pelo fbref.com (2021), uma página de fácil acesso do histórico de estatísticas no futebol que incluem estatísticas de jogadores, times e ligas de diversos países.

O conjunto de dados original apresentava informações de todos os 532 atletas que jogaram o campeonato inglês na temporada de 2020-2021. Além disso, havia a presença de 24 variáveis. Contudo, apenas 272 jogadores foram mantidos para o estudo levando em consideração atletas das posições de ataque e meio de campo e excluindo goleiros e zagueiros que não possuem como objetivo de jogo fazer gols.

Em relação as variáveis, aquelas referentes as informações mais pessoais dos jogadores foram excluídas como nome, nacionalidade, nome da equipe do atleta, posição em campo, ano de nascimento e minutos jogados dividido por 90, uma vez que tais dados não acrescentariam no objetivo do trabalho.

Foram excluídas também 4 variáveis derivadas do xG, sendo elas gols normais previstos, gols normais previstos por chute, quantidade de gols marcados menos xG e quantidade de gols normais marcados menos xG. Dado que xG já será utilizado, as demais não acrescentariam muita informação na análise.

Dessa forma, permaneceu no conjunto de dados final para o estudo as 14 variáveis descritas a seguir:

1. **Idade:** Idade do jogador;
2. **Gols:** Quantidade de gols marcados por jogador na temporada;
3. **TC:** Total de chutes por jogador não incluindo cobranças de pênaltis;
4. **CaG:** Total de chutes por jogador com direção ao gol (excluindo chutes para fora do gol);
5. **SoT:** Porcentagem de chutes por jogador com direção ao gol (CaG/TC);
6. **Sh:** Total de chutes por 90 minutos;
7. **SooT:** Total de chutes com direção ao gol por 90 minutos;
8. **Gols/TC:** Quantidade de gols marcados por total de chutes;
9. **Gols/CaG:** Quantidade de gols marcados por total de chutes com direção ao gol;
10. **Dist:** Distância média do gol, em jardas, de todas as finalizações;
11. **FK:** Chutes de falta por jogador;
12. **PB:** Pênaltis convertidos por jogador;
13. **PT:** Pênaltis batidos por jogador; e
14. **xG:** Gols previstos por jogador (incluem pênaltis).

## 3.1 Análise Descritiva

Com o banco de dados, primeiramente, foi realizada uma análise descritiva. Utilizando o software estatístico R os seguintes resultados foram obtidos.

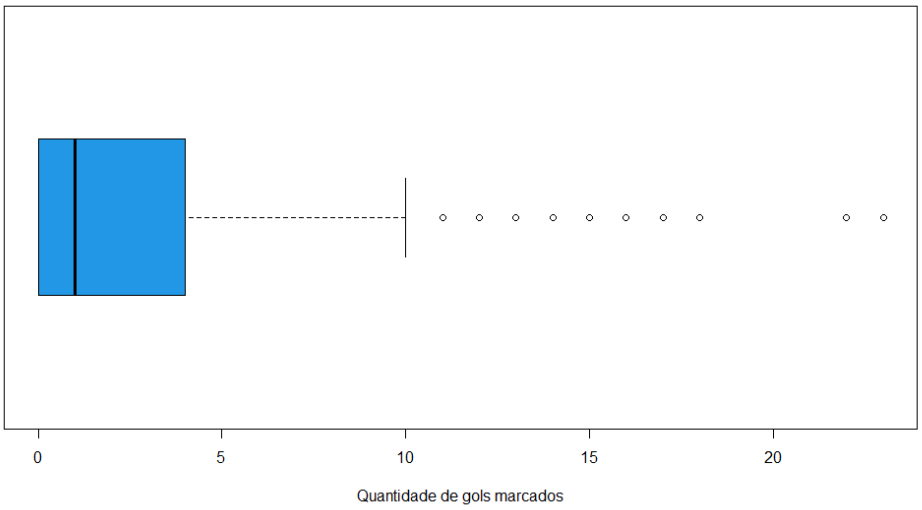


Figura 3.1: Boxplot da quantidade de gols marcados.

Tabela 3.1: Descritiva da quantidade de gols marcados pelos jogadores na liga inglesa.

| Mínimo | 1° Quartil | Mediana | Média | 3° Quartil | Máximo |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 3,015 | 4 | 23 |

A partir da Figura 3.1 e da Tabela 3.1, observamos que os jogadores de linha de frente dos times, isto é, meio campistas e atacantes da Premier League, levando em consideração a temporada 2020-2021, marcam, de uma forma geral, poucos gols. Isto porque 75% dos jogadores analisados marcaram até 4 gols ao longo de todo o campeonato.

Além disso, mais de 25% dos jogadores marcaram zero gols e, tendo isso em vista, a aplicação de um modelo com inflação de zeros poderia ser interessante. Contudo, como os dados serão modelados por um modelo de regressão, tem-se que cada jogador terá sua própria média estimada a partir das variáveis do estudo (não se trata de uma única média) e, portanto, a modelagem será capaz de identificar os jogadores em situação de marcar zero gols.

Por outro lado, podemos identificar alguns pontos extremos na quantidade de gols marcados e os jogadores responsáveis por isto são os chamados artilheiros do campeonato. Destaque para o inglês Harry Kane, jogador do Tottenham, que nessa temporada se sagrou maior goleador com 23 gols marcados ao longo dos 38 jogos.

Apresentado os resultados acima sobre a variável resposta do estudo, o interesse agora é verificar, a princípio de maneira gráfica, quais variáveis preditoras parecem ter uma maior correlação com $Y$. Dessa forma, podemos ter uma ideia inicial de quais fatores

melhor influenciam na performance dos jogadores quanto ao número de gols que eles marcam ao longo do campeonato.

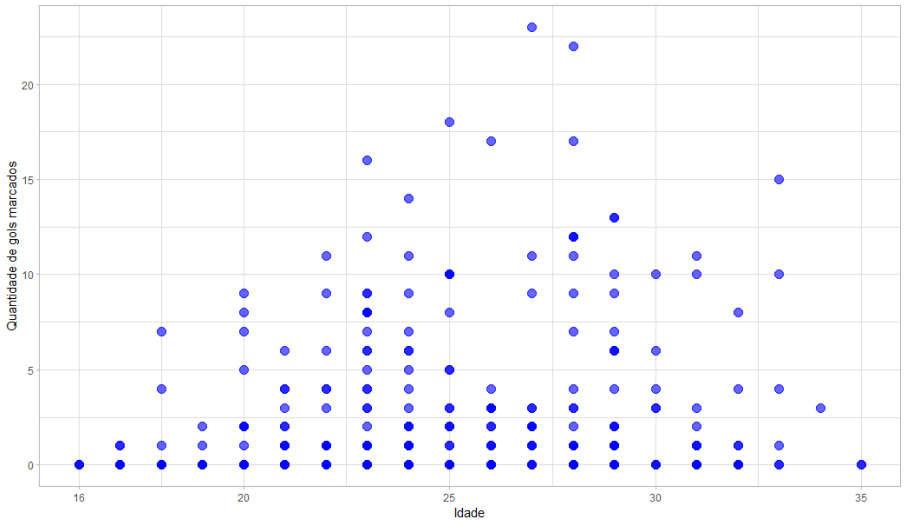


Figura 3.2: Diagrama de dispersão da idade por gols marcados.

A partir do diagrama de dispersão acima, notamos que não há nenhuma relação aparente entre o número de gols que o jogador marca com sua idade. O que descarta a hipótese de que jogadores mais novos com melhores condições físicas ou jogadores mais experientes levam alguma vantagem em relação a seus adversários quando o quesito é tempo de carreira no futebol.

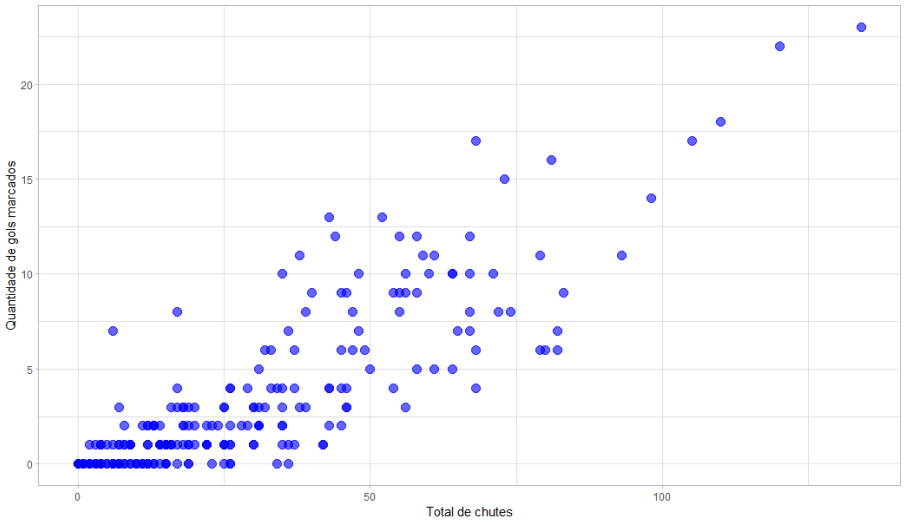


Figura 3.3: Diagrama de dispersão de chutes por gols marcados.

Diferente do caso anterior, a Figura 3.3 mostra uma forte relação linear positiva entre o número total de chutes do jogador e quantidade de gols que ele marca. O que, do ponto de vista prático, parece fazer sentido uma vez que quanto mais o atleta arrisca, maiores são as chances dele obter êxito em suas tentativas. Dessa forma, a variável preditora chutes a gol parece apresentar influência em $Y$.

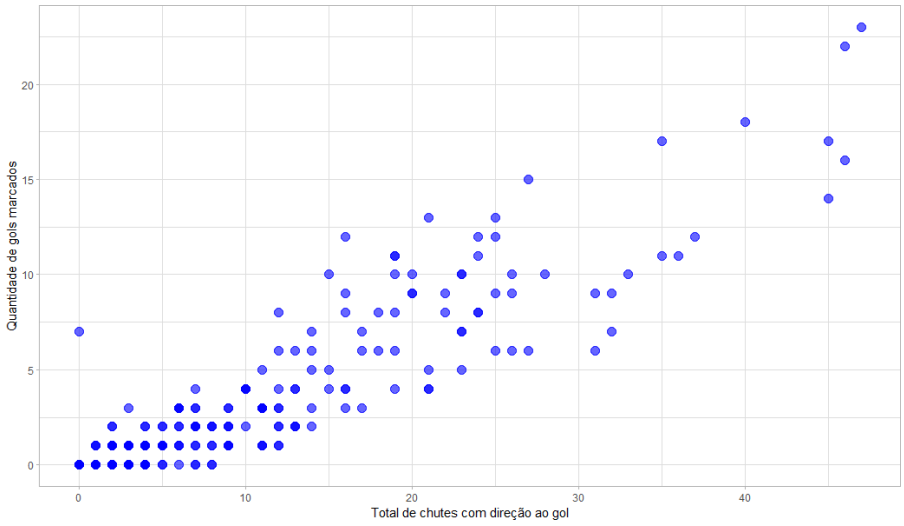


Figura 3.4: Diagrama de dispersão de chutes com direção ao gol por gols marcados.

Da mesma maneira que total de chutes, a variável preditora chutes com direção ao gol apresenta uma forte correlação linear positiva com número de gols marcados. Dessa forma, entendemos que, como visto antes, se a quantidade de gols aumenta conforme a quantidade de chutes também aumenta, faz sentido o fato que quanto mais bolas forem em direção ao gol, e não para fora, as chances de marcar gols continuam altas, se não

maiores. Portanto pontaria no chute parece ser outra variável que apresenta influência em $Y$.

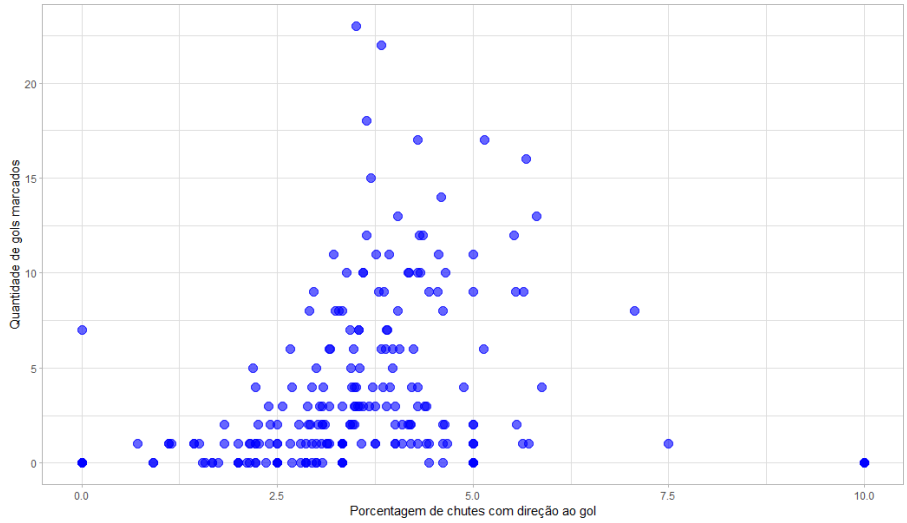


Figura 3.5: Diagrama de dispersão da porcentagem de chutes com direção ao gol por gols marcados.

No caso da Figura 3.5, percebemos que a porcentagem de chutes por jogador com direção ao gol, isto é, total de chutes com direção ao gol dividido pelo número total de chutes, aparenta não ter relação com o número de gols marcados pelos jogadores.

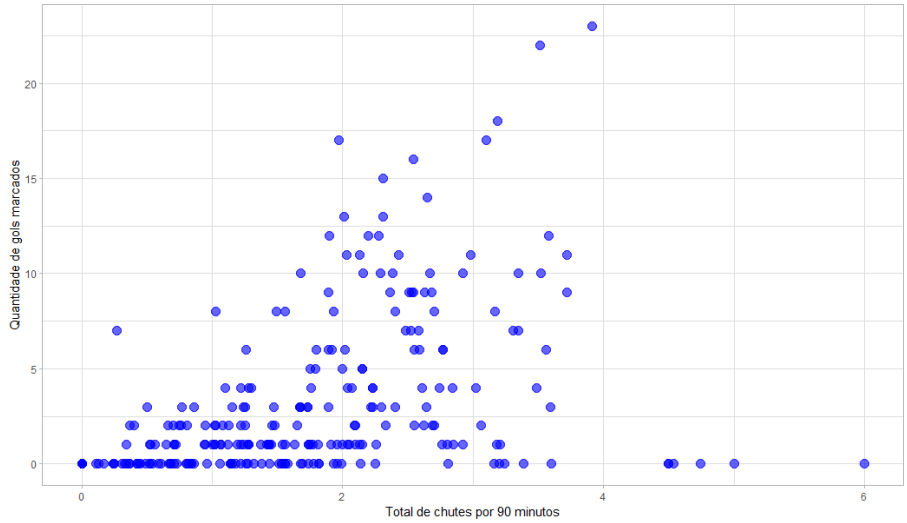


Figura 3.6: Diagrama de dispersão de chutes em 90 minutos por gols marcados.

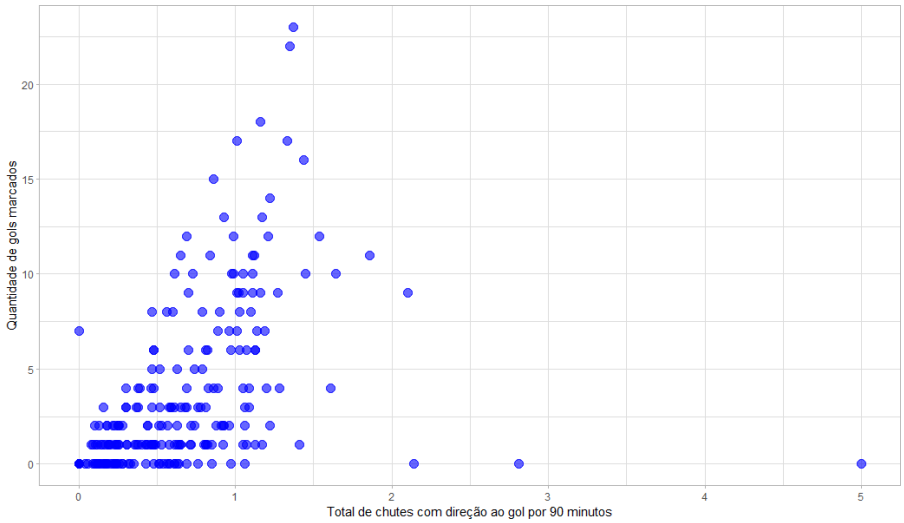


Figura 3.7: Diagrama de dispersão de chutes com direção ao gol em 90 minutos por gols marcados.

Observando as Figuras 3.6 e 3.7 parece que há uma relação linear positiva entre o número total de chutes que o jogador dá (em direção ou não ao gol) dentro dos 90 minutos de jogo levando em consideração as 38 rodadas com a quantidade de gols que este marca. Isto porque, apesar das Figuras apresentarem alguns pontos extremos e a relação em questão ser menos evidente do que outros casos mostrados, conforme o total de chutes (com direção ou não ao gol) em 90 minutos aumenta a quantidade de gols também tende a aumentar.

Ademais, conseguimos observar também que quanto maior o total de chutes com direção ao gol por 90 minutos, maior é a variabilidade de gols marcados.

Em relação aos pontos extremos mais discrepantes com quantidades altas de chutes (com direção ou não ao gol) em 90 minutos e gols marcados iguais a zero são referentes a jogadores que, de fato, obtiveram tais números ao longo do campeonato e, por esse motivo, não serão excluídos do conjunto de dados.

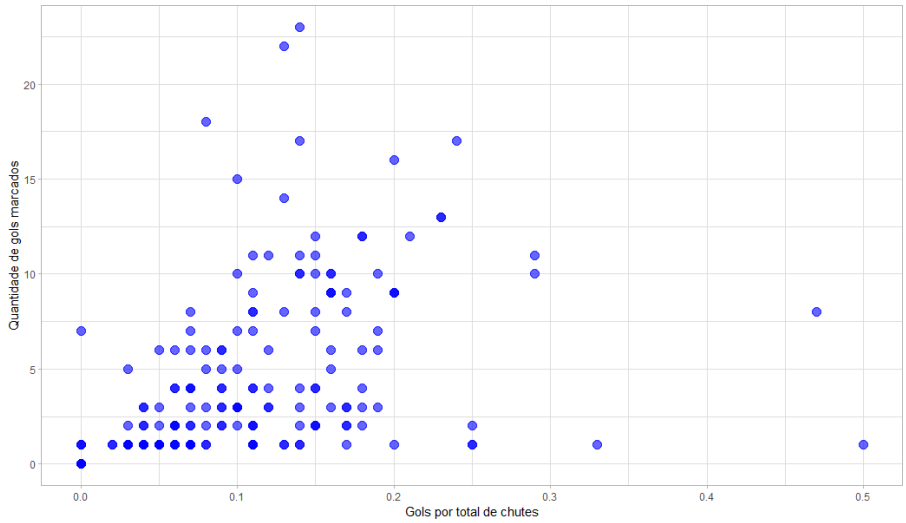


Figura 3.8: Diagrama de dispersão de gols por chutes por gols marcados.

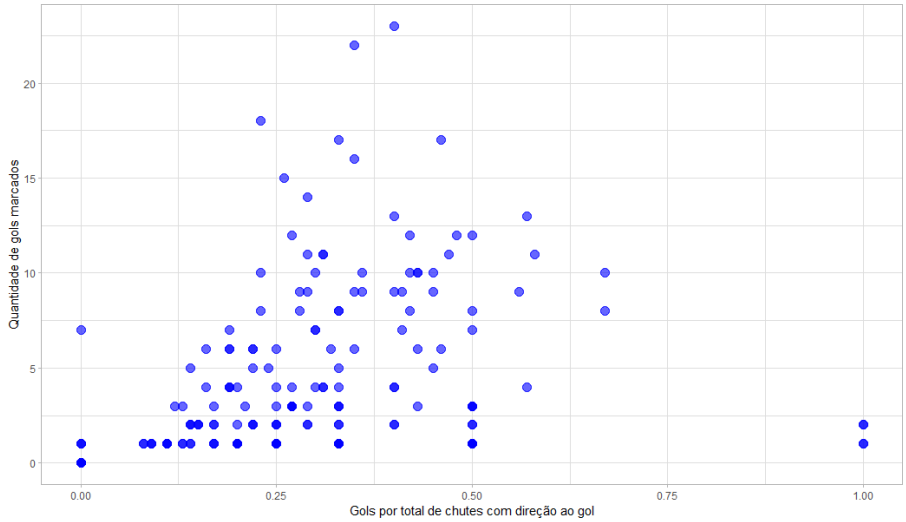


Figura 3.9: Diagrama de dispersão de gols por chutes com direção ao gol por gols marcados.

Já as Figuras 3.8 e 3.9 também parecem mostrar, graficamente, uma não relação entre gols por chutes (em direção ou não ao gol) com a quantidade de gols que o jogador marca ao longo da temporada. Do ponto de vista prático, parece um pouco contraditório dizer que a quantidade de chutes influencia na quantidade de gols marcados mas que gols por chutes não influencia.

Importante ressaltar que, possivelmente, essas duas variáveis sejam descartadas para o ajuste do modelo devido ao fato destas serem função da variável resposta.

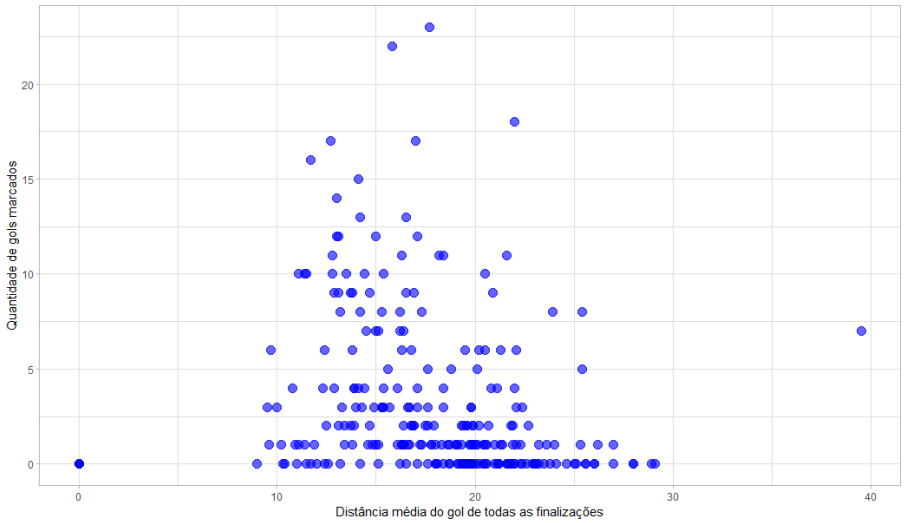


Figura 3.10: Diagrama de dispersão de distância média do gol de todas as finalizações por gols marcados.

Outro fator que parece não influenciar no número de gols que os jogadores marcam, segundo a Figura 3.10, é a distância média do gol de todas as finalizações dos atletas. Podemos entender então que bolas na rede com chutes próximos ou longe do gol não são determinantes para dizer a quantidade final de tentos que jogadores ou equipes irão conseguir fazer nas partidas.

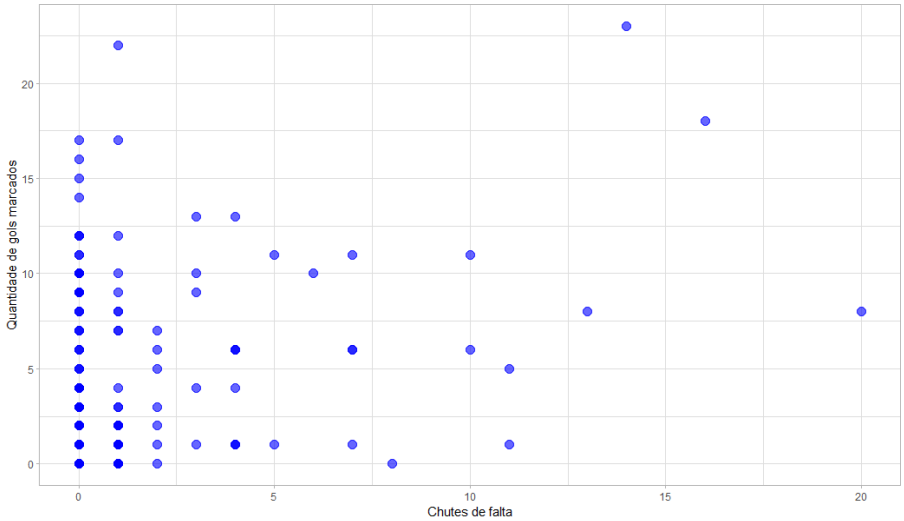


Figura 3.11: Diagrama de dispersão de chutes de falta por gols marcados.

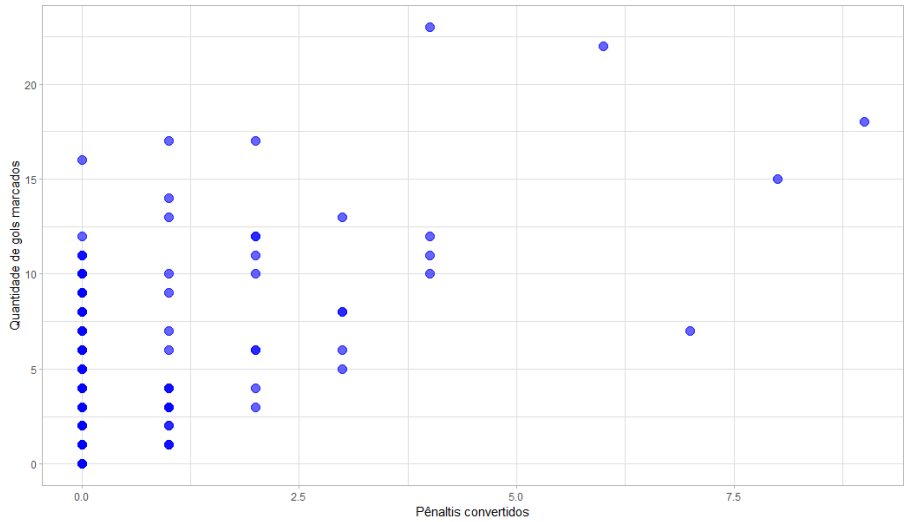


Figura 3.12: Diagrama de dispersão de pênaltis convertidos por gols marcados.

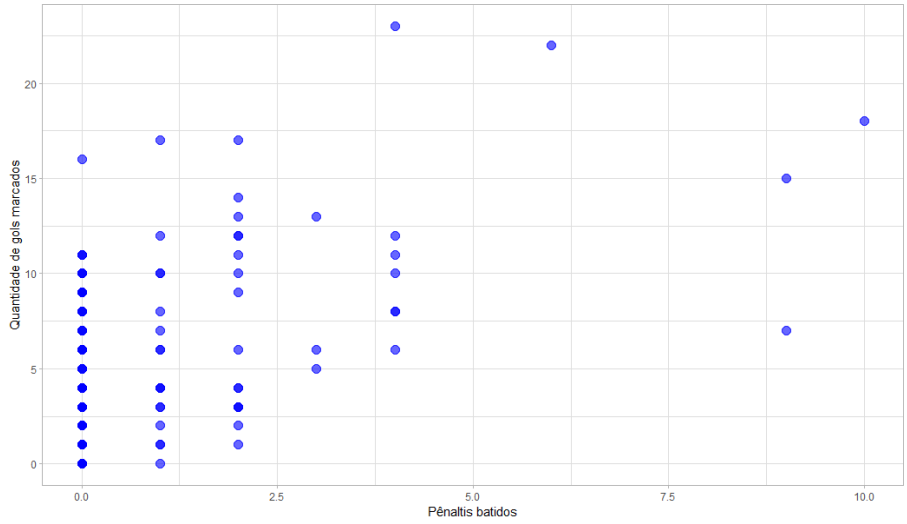


Figura 3.13: Diagrama de dispersão de pênaltis batidos por gols marcados.

Analisando as Figuras 3.11, 3.12 e 3.13 parece que as finalizações com bola parada (faltas e pênaltis) não influenciam na quantidade de gols marcados pelos jogadores. Por outro lado, muitos dos cobradores oficiais de bola parada dos clubes são aqueles que mais chutam ao gol, portanto as oportunidades de bater faltas e pênaltis aumentam ainda mais o número de vezes que estes chutam ao gol e, consequentemente, aumentam a quantidade de gols marcados pelos atletas.

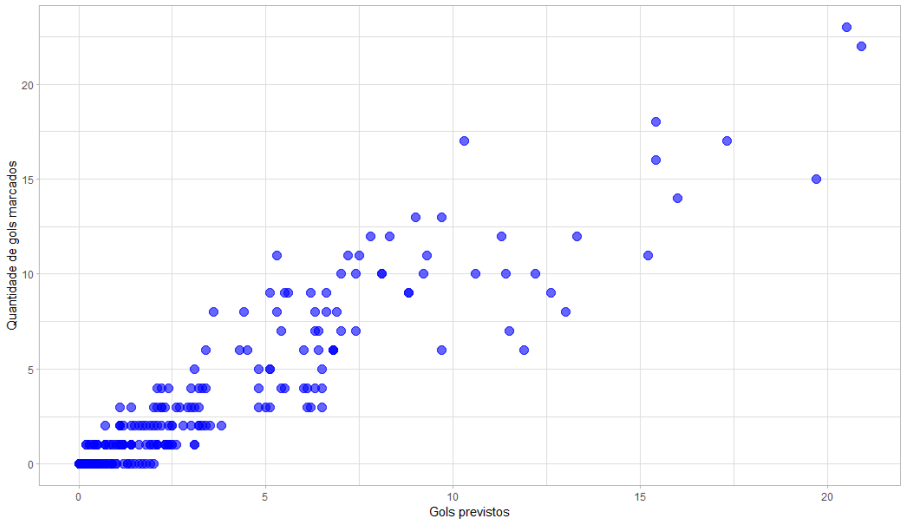


Figura 3.14: Diagrama de dispersão de gols previstos por gols marcados.

Por fim, temos a variável preditora xG que, como explicado no início do trabalho, levando em consideração diversas possibilidades de gols dentro das partidas dado as qualidades de cada atleta, estima a quantidade de gols que os jogadores vão fazer a partir dessas probabilidades analisadas. Segundo a Figura 3.14, essa métrica parece, de fato, influenciar na quantidade de gols que cada jogador realmente faz ao longo da temporada, uma vez que, em geral, quanto maior a quantidade de gols previstos para cada jogador maior é o número de gols que ele marca.

Como observado nas análises acima, tem-se a presença de alguns pontos extremos distribuídos ao longo dos diagramas de dispersões apresentados. Posteriormente no estudo, será realizada uma análise de diagnóstico que, dentre outros aspectos, irá verificar a influência desses pontos nos dados e, a partir dos resultados, serão tratados de maneira adequada.

## 3.2 Análise de correlação entre os jogadores e associação entre as variáveis preditoras

A partir dos dados em estudo, questiona-se a influência que um jogador de um determinado time possui sobre seus companheiros de clube, isto é, será que existe correlação entre a quantidade de gols entre jogadores do mesmo time?

Considerando que a variável resposta de um modelo de regressão binomial negativa é suposta independente entre os indivíduos da amostra, existe a necessidade de analisar a

possibilidade de associação entre jogadores do mesmo clube para garantir que a resposta do estudo, no caso quantidade de gols marcados na temporada, esteja sendo modelada da maneira correta.

Para isto, foi considerado a quantidade de gols dos jogadores analisados como uma série temporal ordenada por time (do melhor classificado no campeonato para o pior) e posição de jogador em campo (primeiro meio campistas de cada time e depois os atacantes), a fim de a partir da construção do gráfico de autocorrelação dessa série, identificar possíveis relações dos jogadores intra grupo. A Figura 3.15 mostra esse gráfico.

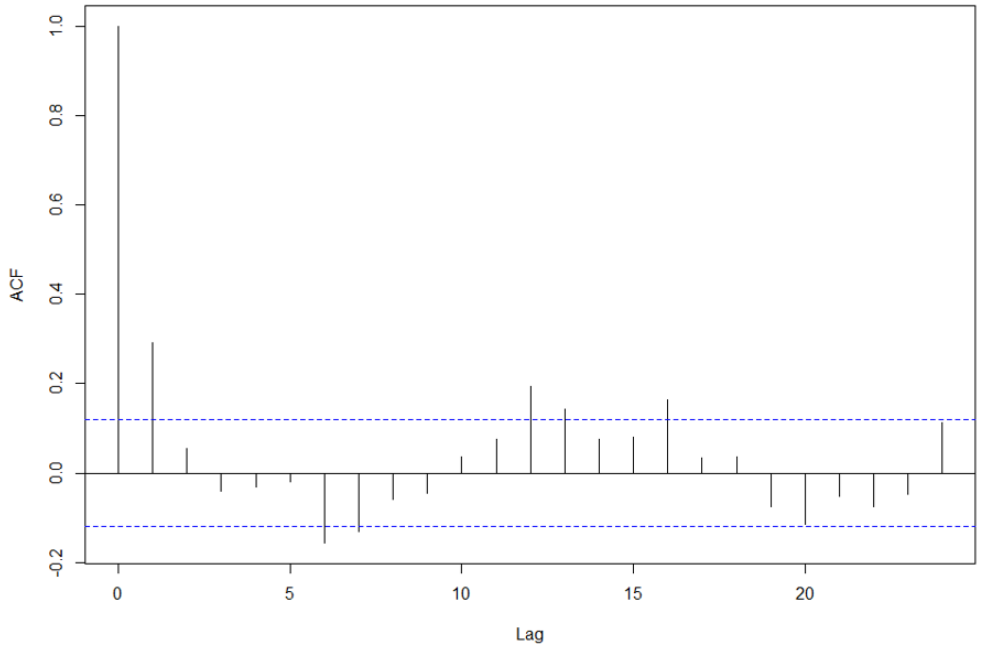


Figura 3.15: Função de Autocorrelação da série.

Observando a Figura 3.15, algumas características chamam atenção. Apesar de em alguns *lags* a autocorrelação ultrapassar as bandas de confiança representada em pontilhado azul, isso acontece de maneira pouco expressiva e apenas a autocorrelação de *lag* 1 possui valor absoluto pouco maior que 0.20.

Portanto, pode-se considerar que a autocorrelação entre jogadores do mesmo time não é substancialmente significativa e, como também será modelado o número médio de gols condicionado às outras características dos jogadores, acreditamos que supor independência entre eles não seja crítico.

Além disso, o gráfico de autocorrelação apresenta ciclos de 12, uma vez que as maiores autocorrelações ocorrem sempre de 6 em 6, intercalando entre uma positiva e outra negativa. Isso ocorre porque, em geral, tem-se de 12 à 17 jogadores ordenados por time

no conjunto de dados.

Outro aspecto importante que pode atrapalhar o ajuste do modelo é a presença de variáveis preditoras altamente correlacionadas entre si e, nesse caso, é importante a seleção daquelas que não apresentem correlações tão fortes para a continuação do estudo.

Nesse sentido, a Figura 3.16 apresenta a matriz de correlação entre as variáveis disponíveis.

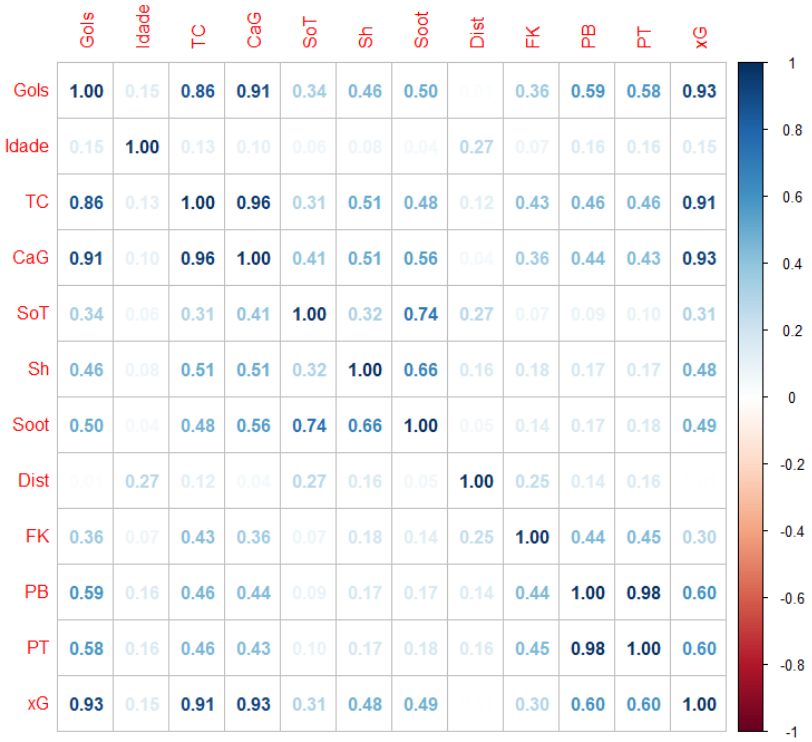


Figura 3.16: Matriz de correlação.

A partir da Figura 3.16, notamos a forte relação entre gols previstos por jogador (xG) com total de chutes por jogador (TC) e total de chutes por jogador com direção ao gol (CaG), ambas correlações maiores que 90%. Possivelmente, o xG de cada jogador é construído com forte influência da quantidade de vezes que o jogador chuta a bola ao gol.

Dessa forma, o xG será mantido no estudo enquanto que as outras citadas serão removidas da análise dado que as informações que estas estão apresentando são praticamente as mesmas do xG. E, além disso, xG será mantido também por apresentar a maior correlação com a variável resposta (93%) em relação as outras duas variáveis (86% e 91%) e por apresentarem uma correlação positiva alta entre elas.

Ademais, vemos uma forte relação entre pênaltis batidos por jogador (PT) com pênaltis convertidos por jogador (PB), com 98% de correlação. Apesar de uma se referir a quan-

tidade de vezes que o jogador bateu pênalti e a outra se referir a quantidade de vezes que a bola de fato entrou no gol, ambas parecem dar a mesma informação. Nesse sentido, a variável pênaltis batidos por jogador (PT) permanecerá no estudo.

Por fim, outra relação que chama a atenção é entre porcentagem de chutes por jogador com direção ao gol (SoT) e total de chutes com direção ao gol por 90 minutos (SooT) com 74% de correlação. De fato, as duas variáveis representam situações bem parecidas no contexto de um jogo de futebol e, para o presente trabalho, será mantido a variável SooT.

Portanto, permanecem na análise as variáveis Idade, Sh, Soot, Dist, FK, PT e xG. Posteriormente, implantado o processo de modelagem aos dados, será levantado o VIF das variáveis em questão para verificar a colinearidade entre elas.

## 3.3 Resultados

### 3.3.1 Modelagem

Como foi visto nas sessões anteriores, o número médio de gols feito por cada jogador ao longo da temporada da Premier League será modelado pela distribuição binomial negativa com a função de ligação logarítmica. Dito isso, ajustamos o seguinte modelo:

$$log(\mu_i) = \beta_0 + \beta_1 \text{Idade}_i + \beta_2 \text{Sh}_i + \beta_3 \text{Soot}_i + \beta_4 \text{Dist}_i + \beta_5 \text{FK}_i + \beta_6 \text{PT}_i + \beta_7 \text{xG}_i. \quad (3.1)$$

Ainda para esse modelo, é necessário definir o valor para o parâmetro de dispersão $\phi$ para que a distribuição do número de gols mencionada pertença à família exponencial e, dessa forma, a modelagem acima possa ser realizada (veja Capítulo 2).

Para isso, a função `glm.nb()` utilizada para a modelagem acaba por estimar o parâmetro $\phi$ a partir de um estimador de momentos após um ajuste inicial usando um MLG Poisson com um determinado número de iterações controlado pelo parâmetro `maxit` da função.

Realizando o processo, foi observado que, a partir de um determinado número de iterações, o valor de $\phi$ fica constante e, mesmo antes de alcançar esse número, o valor permanece muito parecido independente da quantidade de iterações. Nesse sentido, o parâmetro $\phi$ recebeu o valor fixo igual a 3,732.

Dessa forma, levando em consideração o teste de razão de verossimilhança (TRV) e adicionando os termos no modelo de maneira sequencial, temos os resultados mostrados na Tabela 3.2.

Tabela 3.2: ANODEV.

| | GL | Deviance | Pr(>Chi) |
|---|---|---|---|
| **Idade** | 1 | 22,180 | <0,001 |
| **Sh** | 1 | 210,705 | <0,001 |
| **Soot** | 1 | 72,987 | <0,001 |
| **Dist** | 1 | 1,825 | 0,176 |
| **FK** | 1 | 20,059 | <0,001 |
| **PT** | 1 | 49,985 | <0,001 |
| **xG** | 1 | 149,428 | <0,001 |

Ao nível de significância de 5%, temos evidências que, exceto pela distância média do gol de todas as finalizações realizadas pelos atletas, as demais variáveis presentes no modelo são significativas para a quantidade de gols que o mesmo faz ao longo da temporada. Isto é, idade dos jogadores, total de chutes por 90 minutos, total de chutes com direção ao gol por 90 minutos, chutes de falta por jogador, pênaltis batidos por jogador e gols previstos por jogador são fatores importantes para o estudo.

Ademais, o teste de Wald também foi realizado adicionando os termos no modelo de maneira conjunta, isto é, avaliando a significância de uma variável quando o modelo contém todas as demais variáveis. Os resultados se encontram na Tabela 3.3.

Tabela 3.3: Teste de Wald.

| | Estimativa | Erro padrão | $Pr(>\|z\|)$ |
|---|---|---|---|
| **Intercepto** | -1,313 | 0,439 | 0,002 |
| **Idade** | 0,008 | 0,014 | 0,566 |
| **Sh** | -0,054 | 0,089 | 0,546 |
| **Soot** | 0,639 | 0,133 | <0,001 |
| **Dist** | 0,037 | 0,012 | 0,003 |
| **FK** | 0,034 | 0,019 | 0,067 |
| **PT** | -0,095 | 0,043 | 0,027 |
| **xG** | 0,236 | 0,016 | <0,001 |

Diferente do primeiro caso, ao nível de significância de 5%, temos evidências de que apenas total de chutes com direção ao gol por 90 minutos, distância média do gol das finalizações, pênaltis batidos por jogador e gols previstos por jogador são fatores importantes para predizer a quantidade de gols que o atleta faz ao longo da temporada. Além de que a variável distância média do gol das finalizações não tinha apresentado resultado significativo no teste anterior.

Dessa forma, para melhor selecionar as variáveis finais do modelo, foi realizado uma seleção de variáveis tanto pelo método *Stepwise* quanto pelo lasso. Os resultados são apresentados nas Tabelas 3.4 e 3.5.

Tabela 3.4: Método *Stepwise*.

| **Passos** | **Variáveis no modelo** | **AIC** |
|---|---|---|
| **1** | Idade, Sh, Soot, Dist, FK, PT, xG | 927,19 |
| **2** | Idade, Soot, Dist, FK, PT, xG | 925,51 |
| **3** | Soot, Dist, FK, PT, xG | 923,85 |

A partir da Tabela 3.4, observamos que, inicialmente, o método começou com todas as variáveis no modelo e, nos dois passos seguintes, foram retiradas as variáveis idade do jogador e total de chutes por 90 minutos. Considerando o AIC de cada modelo, tem-se que, de fato, não há uma perda significativa retirando-as.

Por outro lado, conduzindo o método lasso pela função "glmnet"e utilizando "cv.glmnet"para escolha do melhor valor de $\lambda$ na validação cruzada, o resultado obtido é apresentado na Tabela 3.5.

Tabela 3.5: Método lasso.

| | **Estimativa** |
|---|---|
| **Intercepto** | 0,567 |
| **Idade** | 0,000 |
| **Sh** | 0,000 |
| **Soot** | 0,506 |
| **Dist** | 0,012 |
| **FK** | 0,022 |
| **PT** | 0,000 |
| **xG** | 0,201 |

Portanto, das 7 variáveis no modelo, o método lasso selecionou 4, sendo elas total de chutes com direção ao gol por 90 minutos, distância média do gol de todas as finalizações, chutes de falta e gols previstos.

Assim como no *Stepwise*, idade e total de chutes por 90 minutos não foram selecionadas e, por conta disso, essas duas variáveis foram retiradas do modelo final. Além disso, o método lasso, diferente do primeiro caso, descartou a variável pênaltis batidos por jogador.

Por fim, o teste de razão de verossimilhança foi novamente realizado tanto para o modelo com as 5 variáveis selecionadas no *Stepwise* quanto para o modelo com as 4 variáveis selecionadas no lasso e, para ambos os casos, foram adicionadas as variáveis de maneira sequencial e, da mesma forma que no primeiro caso, foi considerado um parâmetro $\phi$ com valor igual a 3,526. Os resultados são apresentados abaixo.

Tabela 3.6: ANODEV Modelo *Stepwise*.

| | **GL** | **Deviance** | **Pr(>Chi)** |
|---|---|---|---|
| **Soot** | 1 | 285,964 | <0,001 |
| **Dist** | 1 | 4,067 | 0,043 |
| **FK** | 1 | 21,092 | <0,001 |
| **PT** | 1 | 58,390 | <0,001 |
| **xG** | 1 | 157,228 | <0,001 |

Portanto, ao nível de significância de 5%, temos evidências que o total de chutes com direção ao gol por 90 minutos, distância média do gol de todas as finalizações realizadas pelos atletas, chutes de falta por jogador, pênaltis batidos por jogador e gols previstos por jogador são significativos para a quantidade de gols que o mesmo faz ao longo da temporada.

Tabela 3.7: ANODEV Modelo lasso.

| | **GL** | **Deviance** | **Pr(>Chi)** |
|---|---|---|---|
| **Soot** | 1 | 282,605 | <0,001 |
| **Dist** | 1 | 4,223 | 0,039 |
| **FK** | 1 | 20,229 | <0,001 |
| **xG** | 1 | 207,363 | <0,001 |

Com exceção da variável pênaltis batidos por jogador que não está presente no modelo

lasso, as demais, assim como no primeiro caso, são significativas para a quantidade de gols que o atleta faz ao longo da temporada.

A seguir, será realizada uma análise de diagnóstico para cada modelo apresentado para avaliar qual deles é o mais adequado.

### 3.3.2 Análise de diagnóstico - modelo *Stepwise*

- **Ponto de Alavanca e pontos influentes**

Como observado na análise descritiva, o conjunto de dados apresenta alguns pontos extremos. Dessa forma, o diagnóstico inicial que será realizado no modelo com as 5 variáveis provenientes do *Stepwise* é se esses valores são pontos de alavanca (observações com alto valor na variável resposta e alto valor nas variáveis preditoras) ou pontos influentes (observações que podem alterar de forma significativa o resultado do modelo).

Uma forma de identificar possíveis pontos de alavanca é a partir dos valores da diagonal da matriz $H = \hat{W}^{\frac{1}{2}} X (X' \hat{W} X)^{-1} X' \hat{W}^{\frac{1}{2}}$, também conhecida como matriz de projeção, e analisar as observações com altos valores de $h_{ii}$.

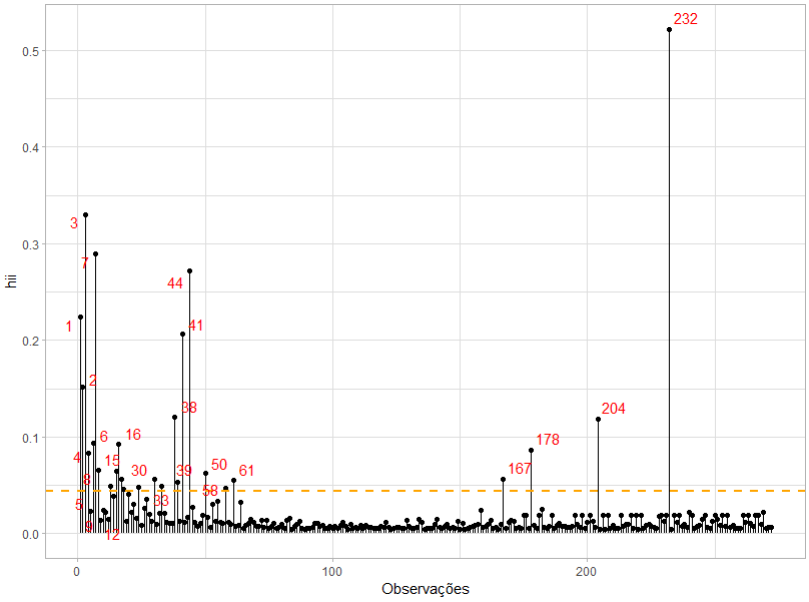


Figura 3.17: Valores da diagonal da matriz de projeção para modelo *Stepwise*.

A Figura 3.17 nos mostra os valores de $h_{ii}$ do modelo com variáveis selecionadas pelo *Stepwise* para cada uma das 252 observações presentes no conjunto de dados. A linha tracejada é o limitante e todas as observações com $h_{ii}$ maior que ela são possíveis pontos de alavanca. Tendo isso em vista, foram 19 observações que se destacaram, sendo elas: 1, 2, 3, 4, 6, 7, 15, 16, 30, 38, 39, 41, 44, 50, 61, 167, 178, 204 e, principalmente, a 232. Portanto, esses pontos são possíveis pontos de alavanca, contudo, também podem ser pontos influentes.

Nesse sentido, uma métrica utilizada para identificar observações influentes é a distância de Cook. A distância de Cook mede a influência da observação $i$ sobre todos $n$ valores ajustados. Essa medida de influência é definida por:

$$D_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(i)})' X' X (\hat{\beta} - \hat{\beta}_{(i)}), \tag{3.2}$$

em que $\hat{\beta}_{(i)}$ é a estimativa do vetor dos coeficientes de regressão desconsiderando a observação $i$ da estimação.

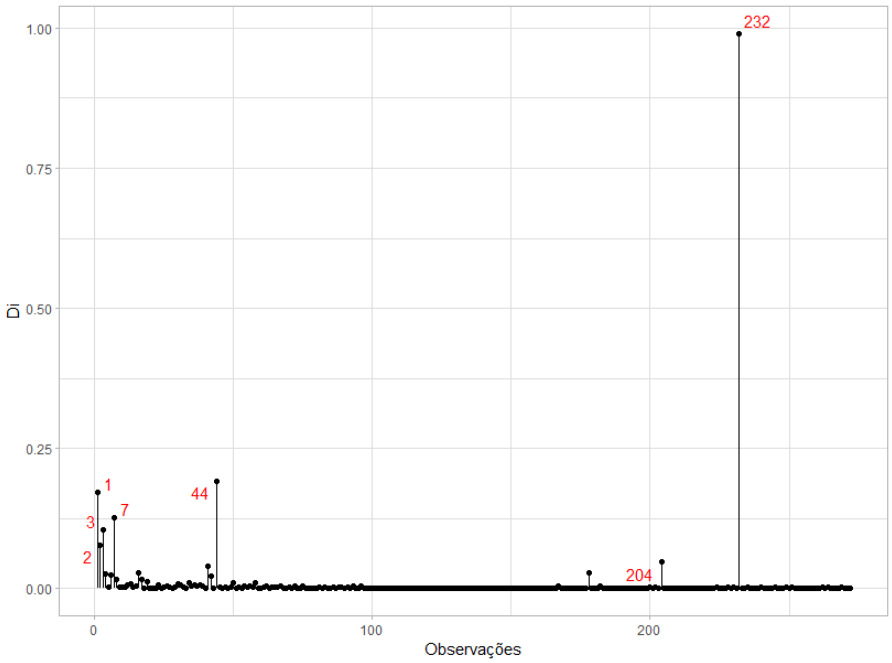


Figura 3.18: Distância de cook para cada observação para modelo *Stepwise*.

A Figura 3.18 nos mostra os valores de $D_i$ para cada uma das 252 observações. Vemos que, novamente, as observações 1, 2, 3, 7, 44, 204 e 232 se destacam assim como nos valores de $h_{ii}$. Dessa forma, analisaremos melhor essas 7 observações que são candidatas a pontos influentes.

Tabela 3.8: Valores das variáveis dos possíveis pontos influentes.

| **Observação** | **Gols** | **Soot** | **Dist** | **FK** | **PT** | **xG** |
|---|---|---|---|---|---|---|
| **1** | 23 | 1.37 | 17.7 | 14 | 4 | 20.5 |
| **2** | 22 | 1.35 | 15.8 | 1 | 6 | 20.9 |
| **3** | 18 | 1.16 | 22 | 16 | 10 | 15.4 |
| **7** | 15 | 0.86 | 14.1 | 0 | 9 | 19.7 |
| **44** | 7 | 0 | 39.5 | 0 | 9 | 7 |
| **204** | 0 | 2.81 | 13 | 0 | 0 | 0 |
| **232** | 0 | 5 | 12.6 | 0 | 0 | 0 |
| **Média** | 3.01 | 0.55 | 16.31 | 0.94 | 0.44 | 3.09 |

De maneira geral, as observações 1, 2, 3, 7 e 44 apresentam número de gols marcados acima da média geral e pelo menos 3 de 5 valores das variáveis preditoras também acima da média. Portanto, se tratam de pontos de alavanca e não influentes.

Por outro lado, as observações 204 e 232, apresentam número de gols marcados igual a zero e a maioria dos valores das variáveis preditoras abaixo da média geral. Nesse caso, temos a presença de duas observações influentes no conjunto de dados.

O restante da análise de diagnóstico do modelo proveniente do método *Stepwise* que será apresentado a seguir já desconsiderou os dois pontos influentes e considera o modelo reestimado sem essas observações.

- **Suposição de adequação da Função de Ligação**

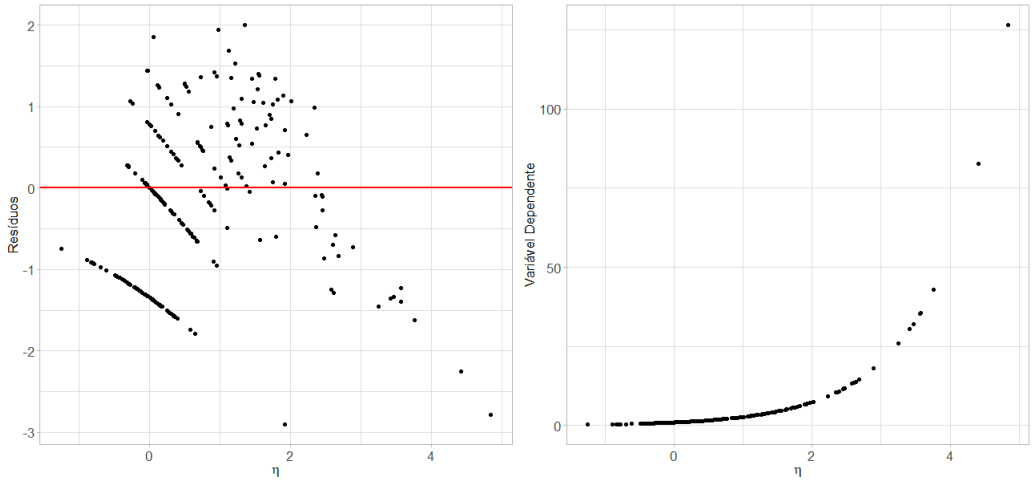


Figura 3.19: Preditor linear versus resíduos e variável dependente ajustada para modelo *Stepwise*.

A partir da Figura 3.19 percebemos que a função de ligação utilizada, no caso a logarítmica, parece não se adequar bem aos dados, uma vez que os pontos no gráfico da esquerda não estão distribuídos de maneira aleatória ao redor de 0 e o pontos do gráfico da direita não apresentam uma relação linear crescente. Algumas transformações foram aplicadas nas variáveis preditoras para tentar corrigir esse comportamento como, por exemplo, transformação quadrática, exponencial e raiz quadrada, mas nenhuma opção melhorou o resultado.

Posteriormente, a mesma análise será feita com o modelo lasso a critério de comparação.

- **Multicolinearidade**

Para verificar a suposição de não colinearidade entre as variáveis preditoras do estudo, foi também analisada a medida do Fator Inflação da Variável $(VIF_j)$ para a $j$-ésima variável preditora que é calculado por:

$$(VIF_j) = \frac{1}{1-R_j^2},$$

em que $R_j^2$ é o coeficiente de determinação múltiplo quando $X_j$ é ajustado por um modelo de regressão linear múltipla em função das demais variáveis preditoras.

Dessa forma, quando $R_j^2$ é próximo de 1 e $1 - R_j^2$ é próximo de zero, maior a correlação linear entre a variável $j$ e as outras variáveis, logo maior é o valor do $VIF$. Por

definição, as variáveis com valores maiores que 10 devem ser avaliadas por haver indícios de colinearidade com outras variáveis presentes no estudo.

A partir disso, os resultados dos VIFs para as variáveis do estudo foram:

Tabela 3.9: $VIF$ de cada variável preditora para modelo *Stepwise*.

| $VIF_1$ | $VIF_2$ | $VIF_3$ | $VIF_4$ | $VIF_5$ |
|---|---|---|---|---|
| 1.546 | 1.424 | 1.423 | 2.193 | 2.728 |

Portanto, obtivemos valores $VIF$ baixos e, dessa forma, concluímos que as variáveis do modelo não são relacionadas entre si.

- **Envelope**

Com o objetivo de verificar se a distribuição utilizada para a variável resposta do modelo, número de gols feitos por jogador ao longo da temporada, foi escolhida adequadamente, foi construído o gráfico envelope para os resíduos apresentado na Figura 3.23.

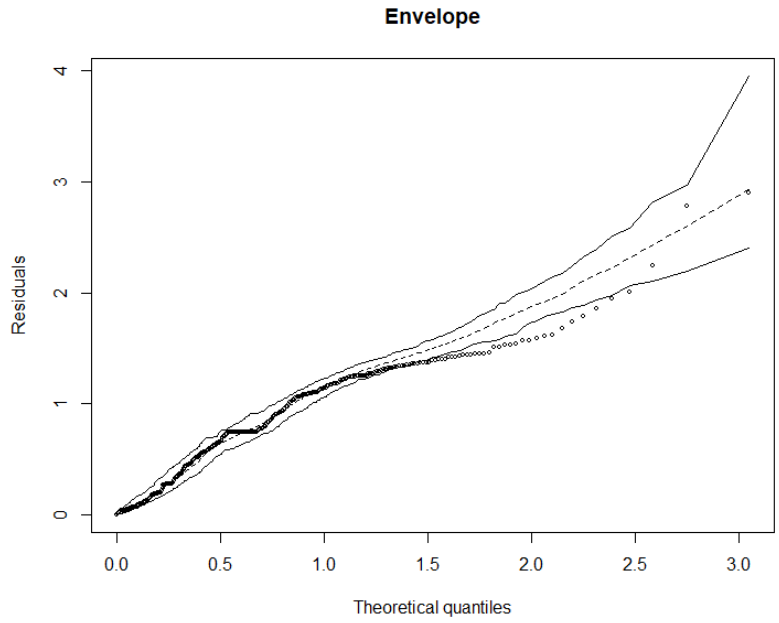


Figura 3.20: Envelope com 14,81% dos pontos para fora para modelo *Stepwise*.

A partir do resultado acima, temos que 14,81% dos pontos (40 de 270 no total) estão fora do envelope, ou seja, existe um número considerável de pontos fora do envelope o que sugere que a distribuição binomial negativa não se adequa de maneira perfeita aos dados. Contudo, ainda podemos considerar um resultado razoável dado que mais de 85% dos pontos estão dentro do envelope.

### 3.3.3 Análise de diagnóstico - modelo lasso

Da mesma forma que o modelo anterior, realizamos a análise de diagnóstico para o modelo com variáveis selecionadas pelo lasso.

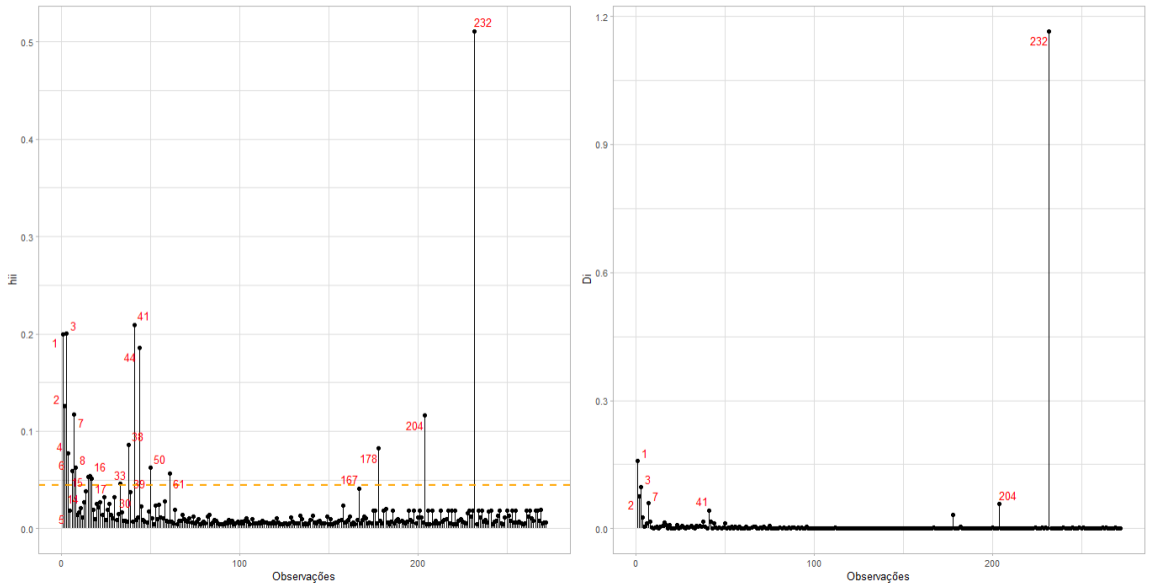


Figura 3.21: Valores da diagonal da matriz de projeção e distância de cook para modelo lasso.

Observando a Figura 3.21 verificamos novamente que a observação 232 e 204 são pontos influentes e, por conta disso, foram desconsideradas do restante da análise e o modelo foi reestimado.

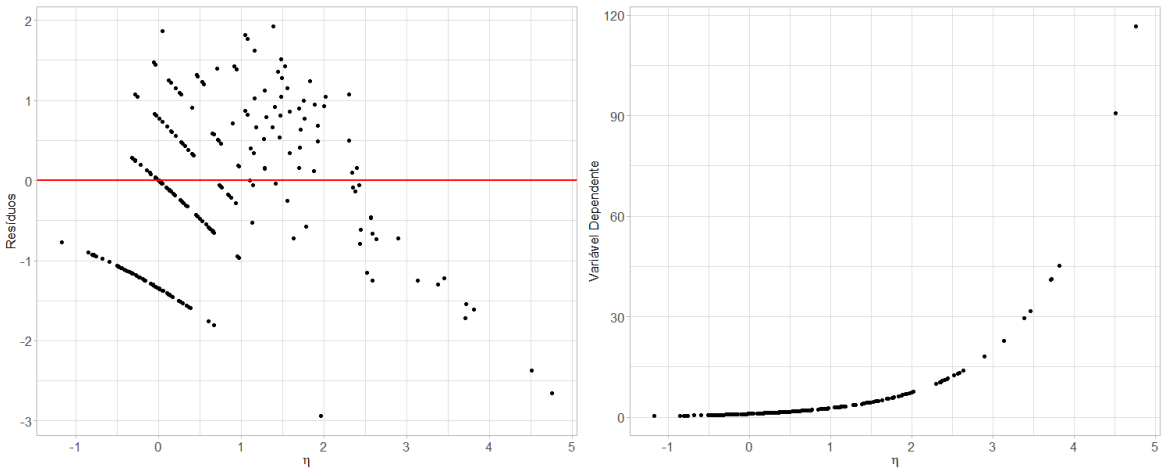


Figura 3.22: Preditor linear versus resíduos e variável dependente ajustada para modelo lasso.

Novamente, pela Figura 3.22 parece que a função de ligação assumida não é adequada

aos dados modelados.

A Tabela 3.10 apresenta os valores de $VIF$ para as 4 covariáveis desse modelo. Portanto, também como no caso anterior, obtivemos valores $VIF$s baixos e, dessa forma, concluímos que as variáveis do modelo não são relacionadas entre si.

Tabela 3.10: $VIF$ de cada variável preditora para modelo lasso.

| $VIF_1$ | $VIF_2$ | $VIF_3$ | $VIF_4$ |
|---|---|---|---|
| 1.442 | 1.291 | 1.288 | 1.587 |

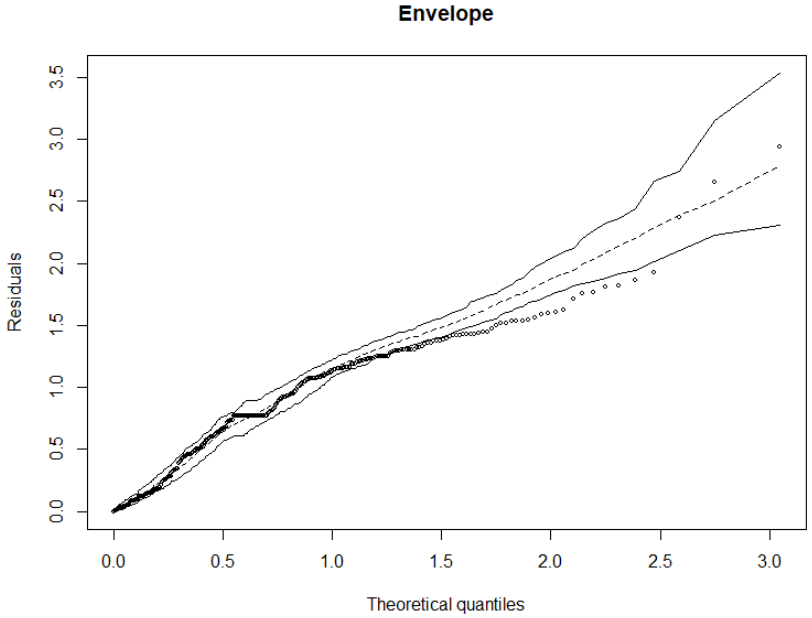


Figura 3.23: Envelope com 19,26% dos pontos para fora para modelo lasso.

Pela Figura 3.23, temos que 19,26% dos pontos (52 de 270 no total) estão fora do envelope. Resultado este ligeiramente inferior ao obtido no modelo *Stepwise*.

Portanto, dado que os dois modelos apresentaram resultados bastante semelhantes quanto à presença de pontos de alavanca e influentes, inadequação da função de ligação e valores $VIF$s baixos, o modelo com a presença da variável pênaltis batidos será mantido para o restante do estudo dado a pequena vantagem observada na adequação da distribuição utilizada para a variável resposta.

### 3.3.4 Modelos alternativos

Como vimos que ambos os modelos estimados na seção anterior não se mostraram tão adequados aos dados, modelos alternativos foram ajustados. Outros valores de $\phi$ foram fixados, já que esse parâmetro é estimado e fixado em um passo anterior à estimação do modelo final. Contudo, os resultados se mantiveram bastante parecidos.

Outra alternativa foi ajustar o modelo *Stepwise* considerando interação de segunda ordem entre as variáveis para avaliar se, com essas inclusões no preditor linear, as análises de diagnósticos melhoravam. Contudo, os resultados também permaneceram bastante semelhantes.

A troca da função de ligação também se tornou uma opção. Utilizamos a função de ligação raiz quadrada com as variáveis provenientes do *Stepwise* e os resultados são apresentados a seguir.

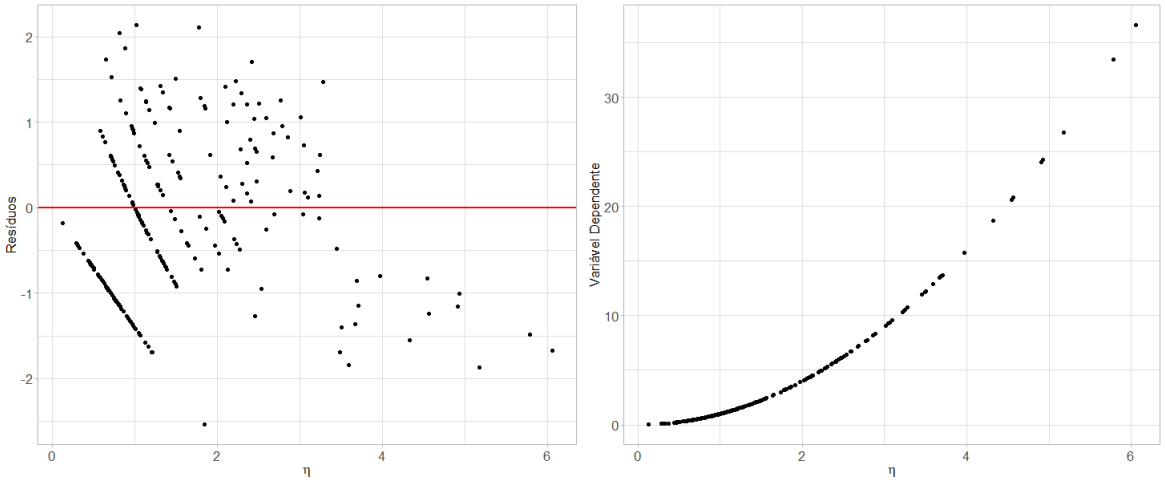


Figura 3.24: Preditor linear versus resíduos e variável dependente ajustada para modelo com distribuição binomial negativa considerando função de ligação raiz quadrada.

Na Figura 3.24 percebemos então uma melhor distribuição dos pontos ao redor do zero no gráfico à esquerda e um comportamento mais linear crescente no gráfico à direita. O que indica que essa função de ligação é melhor do que a anterior.

Contudo, pela Figura 3.25 que mostra o envelope para os resíduos do modelo estimado, observamos que 71,48% dos pontos (193 de 270 no total) estão fora do envelope. Apesar de uma grande proporção dos pontos estar fora do envelope, observamos que eles se distanciam por muito pouco do limite inferior do envelope e apresentam, no geral, o formato mais próximo de uma reta entre os já obtidos. Os resíduos desse modelo também são os menores em valores absolutos, o que evidencia um ajuste mais preciso aos dados.

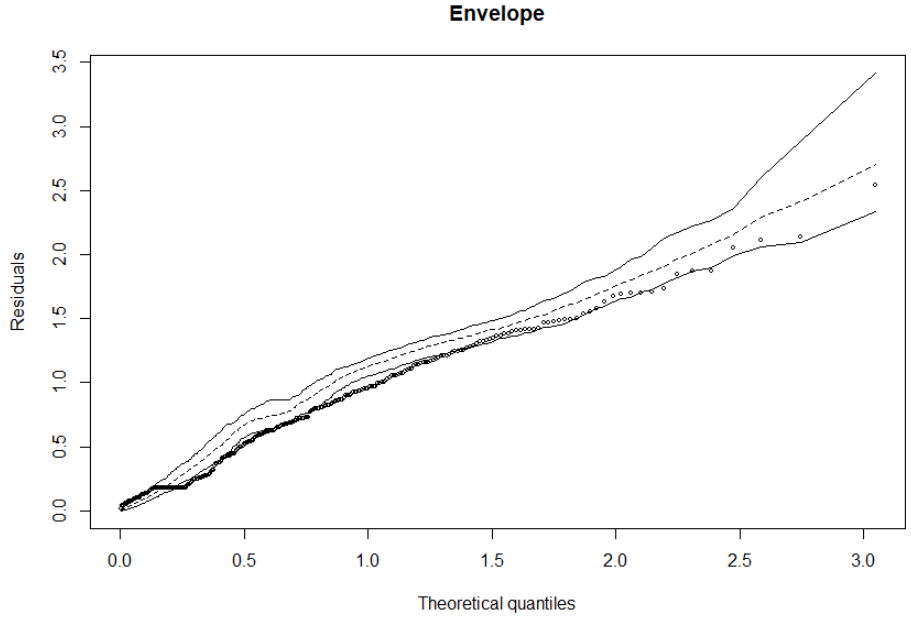


Figura 3.25: Envelope com 71,48% dos pontos para fora para modelo com distribuição binomial negativa considerando função de ligação raiz quadrada.

Levando tudo isso em consideração, um novo modelo foi ajustado considerando as 5 variáveis definidas acima, sem a presença das duas observações influentes, função de ligação logarítmica e com a distribuição da variável resposta, número de gols feitos ao longo da temporada, sendo a distribuição Poisson. Abaixo são apresentados seus resultados.

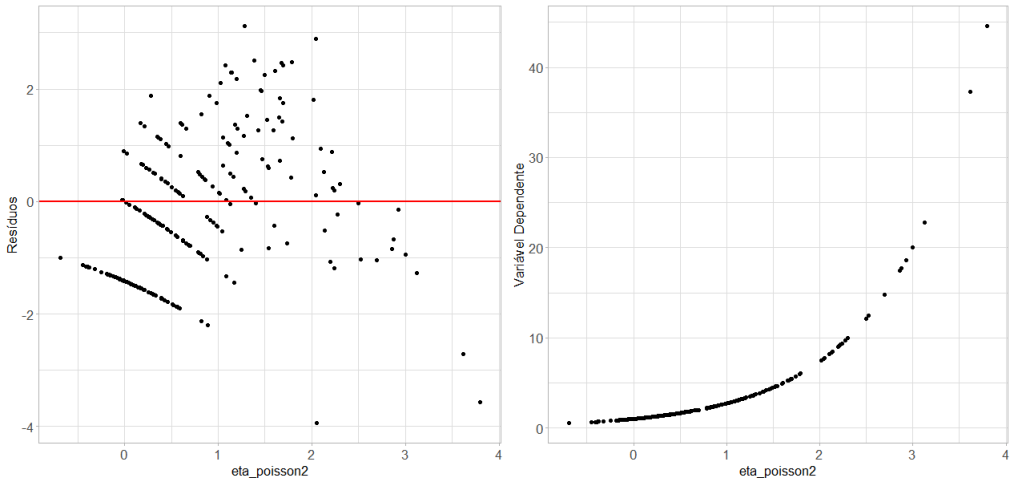


Figura 3.26: Preditor linear versus resíduos e variável dependente ajustada para modelo com distribuição Poisson considerando função de ligação logarítmica.

Os resultados da Figura 3.26 mostram que, mesmo considerando uma distribuição diferente para a variável resposta, a função de ligação logarítmica continua não sendo

adequada aos dados modelados.

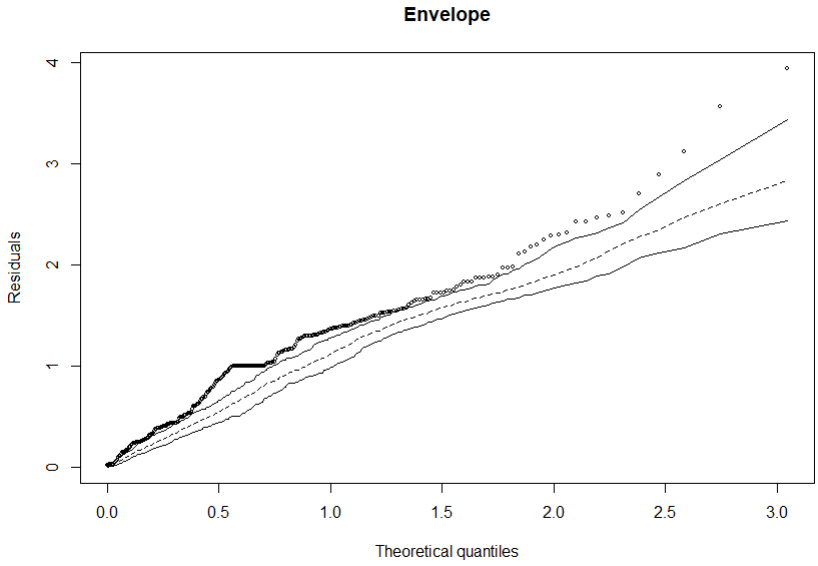


Figura 3.27: Envelope com 94,81% dos pontos para fora para modelo com distribuição Poisson considerando função de ligação logarítmica.

A partir da Figura 3.27, vemos que 94,81% dos pontos estão fora do envelope. Resultado este ainda mais inferior ao apresentado pelos resíduos do modelo binomial negativo utilizando função de ligação raiz quadrada.

Dessa forma, outro modelo também foi ajustado considerando agora as mesmas 5 variáveis, sem a presença das duas observações influentes, função de ligação raiz quadrada e com a distribuição da variável resposta sendo, novamente, a distribuição Poisson. Abaixo são apresentados seus resultados.

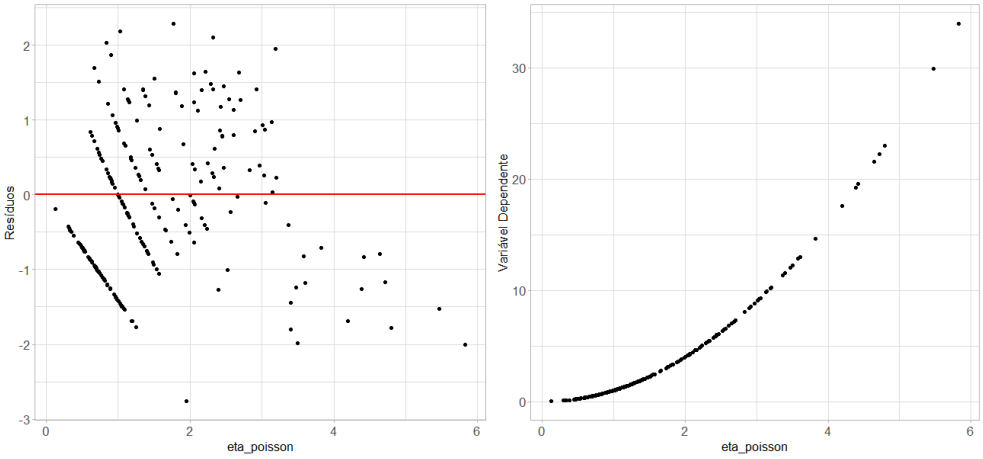


Figura 3.28: Preditor linear versus resíduos e variável dependente ajustada para modelo com distribuição Poisson considerando função de ligação raiz quadrada.

Os resultados da Figura 3.28 são bem parecidos com os obtidos no modelo binomial negativo usando a função de ligação raiz quadrada. Isso evidencia que essa função de ligação parece ser, de fato, mais adequada aos dados do que a função de ligação logarítmica.

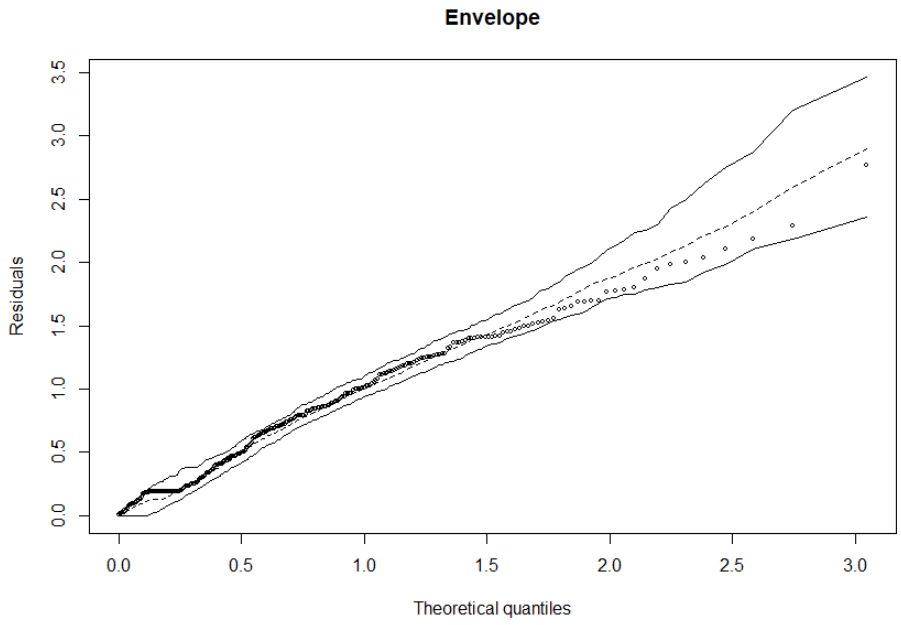


Figura 3.29: Envelope com 0,74% dos pontos para fora para modelo com distribuição Poisson considerando função de ligação raiz quadrada.

Além da análise da função de ligação, quando observamos a Figura 3.29, temos o melhor envelope dentre todos os construídos com apenas 0,74% dos pontos fora do mesmo.

Dessa maneira, pelo gráfico de envelope dos resíduos, o modelo Poisson parece ser mais adequado que o modelo binomial negativo com a função de ligação raiz quadrada. No entanto, considerando as outras análises, os dois modelos são muito parecidos e o modelo binomial negativo apresenta os menores resíduos em valores absolutos.

Como o modelo Poisson é um caso especial e limite do binomial negativo (quando $\phi \to \infty$), os resultados do modelo binomial negativo deveriam ser piores quando o $\phi$ estivesse mal estimado. Dessa maneira, ajustamos outros modelos binomial negativo para diferentes e grandes valores de $\phi$, mas os resultados da análise de diagnóstico não se alteraram.

Outro resultado observado é apresentado na Figura 3.11 a partir do valor de AIC de cada um dos modelos ajustados com a presença das 5 variáveis provenientes do método *Stepwise* e sem as duas observações influentes.

Tabela 3.11: Valores de AIC dos modelos analisados

| | **Distribuição** | |
|---|---|---|
| **Função de ligação** | Binomial negativa | Poisson |
| Logarítmica | 925,9 | 956,9 |
| Raiz quadrada | 776,6 | 777,9 |

Percebemos então que a função de ligação é, assim como analisado anteriormente, determinante para a obtenção de um modelo que melhor se adeque aos dados. Vemos que ambos os modelos com função de ligação logarítmica são significantemente piores em relação aos modelos com função de ligação raiz quadrada quando comparamos seus valores de AIC.

Por outro lado, confirmando também o que foi visto anteriormente, tanto o modelo com distribuição binomial negativa quanto o modelo com distribuição Poisson, ambos considerando a função de ligação raiz quadrada, são os melhores modelos dentre os analisados e muito semelhantes entre si.

Na Tabela 3.12, apresentamos as estimativas dos coeficientes de regressão tanto do modelo Poisson quanto do binomial negativo usando a raiz quadrada como função de ligação e sem as duas observações influentes.

Tabela 3.12: Estimativas dos coeficientes de regressão no modelo Poisson e binomial negativa considerando função de ligação raiz quadrada.

| | **Poisson** | | **Binomial negativo** | |
|---|---|---|---|---|
| | **Estimativa** | **Pr(>Chi)** | **Estimativa** | **Pr(>Chi)** |
| **Soot** | 0,631 | <0,001 | 0,583 | <0,001 |
| **Dist** | 0,016 | <0,001 | 0,016 | <0,001 |
| **FK** | 0,027 | 0,043 | 0,026 | 0,077 |
| **PT** | -0,021 | 0,516 | -0,019 | 0,601 |
| **xG** | 0,206 | <0,001 | 0,223 | <0,001 |

Apesar de verificado anteriormente que o modelo *Stepwise* com as 5 variáveis apresentava uma ligeira vantagem em relação ao modelo lasso com 4 variáveis (excluindo PT), na Tabela 3.12 observamos que essa variável PT, pênaltis batidos por jogador, quando considerada função de ligação raiz quadrada, não é significativa para ambos os modelos pelos testes de significância. Dessa forma, essa variável será descartada sem grandes impactos

na estimativa dos outros coeficientes e, consequentemente, na análise de diagnóstico dos modelos.

Além disso, a variável FK, chutes de falta por jogador, se mostra significativa para o modelo Poisson e não significativa para o modelo binomial negativo. Contudo, considerando um $\alpha = 5\%$, o p-valor dessa variável no segundo modelo é bem próximo ao nível de significância e ela será mantida no modelo.

Por fim, a Tabela 3.13 apresenta as estimativas dos modelos considerando a alteração mencionada.

Tabela 3.13: Estimativa dos coeficientes de regressão do modelo Poisson e binomial negativa considerando função de ligação raiz quadrada sem variável PT.

| | **Poisson** | | **Binomial negativo** | |
|---|---|---|---|---|
| | **Estimativa** | **Pr(>Chi)** | **Estimativa** | **Pr(>Chi)** |
| **Soot** | 0,650 | <0,001 | 0,598 | <0,001 |
| **Dist** | 0,016 | <0,001 | 0,015 | <0,001 |
| **FK** | 0,024 | 0,05 | 0,024 | 0,09 |
| **xG** | 0,201 | <0,001 | 0,218 | <0,001 |

Observamos pelo modelo Poisson e binomial negativo que todas as variáveis dos modelos finais apresentam coeficiente positivo e valor parecido entre eles, ou seja, à medida que o jogador chuta mais em direção ao gol por 90 minutos (Soot), a distância média do gol (Dist), o total de chutes de falta (FK) e o total de gol previstos (xG) aumentam, o número de gols médios no campeonato também é maior.

Finalmente, de acordo com o objetivo do trabalho, observamos que entre as variáveis disponíveis para o estudo, as que impactam o número de gol de um jogador no campeonato inglês são: total de chutes com direção ao gol por 90 minutos, a distância média do gol de todas as finalizações, os chutes de falta por jogador e os gols previstos por jogador.

# Capítulo 4

# Conclusão e estudos futuros

Em uma partida de futebol, o principal objetivo do time é marcar mais gols do que o seu adversário para sair vencedor do confronto. Contudo, preparação e treinamento são fatores fundamentais aos jogadores para que consigam alcançar bons resultados dentro de campo. Nesse sentido, o intuito central desse estudo foi buscar fatores que, dado a devida atenção, podem potencializar a capacidade de um jogador em fazer gols ao longo de uma temporada inteira e, para isso, dados do campeonato inglês de futebol de 2020 e 2021 foram utilizados na análise.

Como a quantidade de gols marcados se trata de uma variável de contagem e tal variável pode apresentar uma alta variabilidade, a distribuição binomial negativa foi inicialmente pensada como sendo mais adequada a esses dados do que a Poisson dado sua maior flexibilidade.

Dessa forma, tratamos de um modelo linear generalizado com função de ligação logarítmica, a mais usual para uma regressão binomial negativa. E vimos que, após supor independência entre jogadores de um mesmo time e excluir variáveis altamente correlacionadas entre si, o método de seleção de variáveis *Stepwise* apresentou como fatores importantes na quantidade de gols que o jogador marca ao longo da temporada o total de chutes com direção ao gol por 90 minutos, a distância média do gol de todas as finalizações, os chutes de falta, os pênaltis batidos e os gols previstos por jogador. Por outro lado, o método de seleção lasso selecionou as mesmas variáveis com exceção dos pênaltis batidos por jogador.

Um análise de diagnóstico foi realizada para ambos os modelos para identificar qual deles era melhor e, com isso, apesar do modelo *Stepwise* com 5 variáveis apresentar uma leve superioridade, vimos que modelos com distribuição binomial negativa e função de

ligação logarítmica pareciam não se adequar bem aos dados. Por conta disso, outras alternativas foram testadas e as que resultaram em melhores diagnósticos foram o modelo Poisson considerando função de ligação raiz quadrada e o modelo binomial negativo considerando também função de ligação raiz quadrada.

Esses dois modelos apresentaram coeficientes de regressão estimados e desempenhos muito parecidos. O modelo binomial negativo apresentou os menores valores absolutos de resíduos, enquanto que o modelo Poisson parece ser levemente mais adequado quando vimos o gráfico de envelope para os resíduos.

Contudo, o resultado mais importante foi, ao final do processo, identificar que os melhores modelos ajustados mostraram que o total de chutes com direção ao gol por 90 minutos, a distância média do gol de todas as finalizações, os chutes de falta e os gols previstos por jogador são significativos na explicação da quantidade de gols que o atleta faz ao longo do campeonato inglês. Portanto, o presente estudo traz algumas variáveis que podem ser melhores estudadas e adequadas aos treinamentos do jogador para, dessa forma, existir a possibilidade de potencializar sua capacidade de marcar gols.

Como sugestão a trabalhos futuros com o intuito de incrementar a análise feita ou até mesmo torná-la mais completa, seguem algumas abordagens que podem ser utilizadas:

- Modelos mais flexíveis que contemplem dados inflacionados de zero ou acomodem super ou subdispersão, como a Poisson inflacionada de zero ou a Conway–Maxwell–Poisson (Wikipédia, 2023);
- Modelo MLG misto com o intercepto aleatório que acomode qualquer possível associação entre os jogadores do mesmo time a fim de analisar formalmente a necessidade de considerar correlação entre eles;
- Combinação de variáveis preditoras muito correlacionadas via componentes principais e utilizar as componentes principais relevantes na análise, em vez de excluir algumas delas; e
- Árvores Híbridas com um componente paramétrico e as demais variáveis como variáveis particionadoras possibilitando a modelagem de interações de qualquer ordem e de relações não lineares entre as variáveis.

# Referências Bibliográficas


Caley, M. (2014). What is the best method for predicting football matches? Disponível em: https://cartilagefreecaptain.sbnation.com/2014/3/5/5473358/what-is-the-best-method-for-predicting-football-matches. Acesso em: 18 setembro 2021.

Cameron, A. C. e Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Cordeiro, G. M. e Demétrio, C. G. (2008). *Modelos lineares generalizados e extensões*. Piracicaba: USP.

da Silva, B. M. (2018). Multiple linear regression applied to football/regressao linear multipla aplicada ao futebol. *Revista Brasileira de Futsal e Futebol*, **10**(38), 262–271.

fbref.com (2021). 2020-2021 premier league estatísticas de chutes. Disponível em: https://fbref.com/pt/comps/9/10728/shooting/2020-2021-Premier-League-estatisticas. Acesso em: 05 setembro 2021.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.*. Sage Publications, Inc.

Friedman, J., Hastie, T. e Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, **49**(2), 127–145.

LAW, J. (2020). Cultura das estatísticas aproxima time pequeno de londres da premier league. Disponível em: https://www.uol.com.br/esporte/futebol/ultimas-noticias/2020/06/11/

estatistica-empurra-time-modesto-de-londres-em-sonho-de-ir-a-premier-league.htm. Acesso em: 15 setembro 2021.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, **15**(3), 209–225.

Neter, J., Wasserman, W. e Kutner, M. H. (1989). *Applied linear regression models*. Irwin Homewood, IL.

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

StatsBomb (2022). What are expected goals (xg)? Disponível em: https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/#:~:text=Put%20simply%2C%20Expected%20Goals%20(xG,scale%20between%200%20and%201. Acesso em: 05 fevereiro 2023.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

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

Venables, W. N. e Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Wikipédia (2023). Conway–maxwell–poisson distribution. Disponível em: https://en.wikipedia.org/wiki/Conway%E2%80%93Maxwell%E2%80%93Poisson_distribution. Acesso em: 06 fevereiro 2023.

# Apêndice A

# Códigos

```
########## COMANDOS TCC ###########

## BANCO DE DADOS ##
library(readxl)
Dados <- read_excel("E:/DISCIPLINAS_GRADUAÇÃO/TCC/Dados.xlsx")
View(Dados)

## VARIÁVEIS ##
Gols <- Dados$Gols
var(Gols)
Idade <- Dados$Idade
TC <- Dados$TC
CaG <- Dados$CaG
SoT <- Dados$‘SoT%‘
Sh <- Dados$‘Sh/90‘
Soot <- Dados$‘SoT/90‘
GSh <- Dados$‘G/Sh‘
GSoT <- Dados$‘G/SoT‘
Dist <- Dados$Dist
FK <- Dados$FK
PB <- Dados$PB
PT <- Dados$PT
xG <- Dados$xG
```

```
df <- data.frame(Gols, Idade, TC, CaG, SoT, Sh, Soot, GSh, GSoT, Dist, FK, PB, PT, xG)

## ANÁLISE DESCRITIVA ##

# boxplot Quantidade de gols marcados

ggplot(data = Dados, aes(y = Gols)) +
  geom_boxplot(color = "paleturquoise3",
               fill = "paleturquoise2",
               outlier.color = "dodgerblue") +
  labs(x = "Resíduos")

boxplot(Gols, data = Dados, col = 4,
        xlab="Quantidade de gols marcados",horizontal=TRUE)

summary(Gols)

# Gráfico de dispersão Idade com Gols
library(ggplot2)

(a <- ggplot(Dados, aes(Idade, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    scale_x_continuous(breaks = c(16,20,25,30,35)) +
    labs(x = "Idade", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Total de chutes por jogador com Gols
(b <- ggplot(Dados, aes(TC, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
```

```
    labs(x = "Total de chutes", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Total de chutes com direção ao gol com Gols
(c <- ggplot(Dados, aes(CaG, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Total de chutes com direção ao gol", y =
    "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Porcentagem de chutes com direção ao gol com Gols
(d <- ggplot(Dados, aes(SoT, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Porcentagem de chutes com direção ao
    gol", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Total de chutes por 90 minutos com Gols
(e <- ggplot(Dados, aes(Sh, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Total de chutes por 90 minutos", y =
    "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
```

```
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Total de chutes com direção ao gol por 90 minutos com Gols
(f <- ggplot(Dados, aes(Soot, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Total de chutes com direção ao gol por 90
    minutos", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Gols por total de chutes com Gols
(g <- ggplot(Dados, aes(GSh, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Gols por total de chutes", y =
    "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Gols por total de chutes com direção ao gol com Gols
(h <- ggplot(Dados, aes(GSoT, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Gols por total de chutes com direção ao
    gol", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()
```

```
# Gráfico de dispersão Distância média do gol de todas as finalizações com Gols
(i <- ggplot(Dados, aes(Dist, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Distância média do gol de todas as
    finalizações", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Chutes de falta com Gols
(j <- ggplot(Dados, aes(FK, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Chutes de falta", y = "Quantidade de gols
    marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Pênaltis convertidos com Gols
(k <- ggplot(Dados, aes(PB, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Pênaltis convertidos", y = "Quantidade de
    gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Pênaltis batidos com Gols
(l <- ggplot(Dados, aes(PT, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
```

```
    labs(x = "Pênaltis batidos", y = "Quantidade de
    gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

# Gráfico de dispersão Gols previstos por jogador com Gols
(m <- ggplot(Dados, aes(xG, Gols)) +
    geom_point(alpha = 0.6, color = "blue", size = 3.5, shape = 19) +
    labs(x = "Gols previstos", y = "Quantidade de gols
    marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))) + theme_light()

### Análise de correlação entre os jogadores

## BANCO DE DADOS ##
library(readxl)
Dados2 <- read_excel("E:/DISCIPLINAS_GRADUAÇÃO/TCC/Dados_Análise_C
orrelação_Jogadores.xlsx")
View(Dados2)

#Pacotes Utilizados
library(forecast)
library(stats)
library(astsa)
library(randtests)

## série temporal ##
serie_temporal = ts(Dados2$Gols)
```

```
ts.plot(serie_temporal)

## gráfico da função de auto correlação ##
acf(serie_temporal)

## Correlação entre variáveis
library(corrplot)

names(Dados)[c(6,7,8)] <- c("SoT","Sh","Soot")
Dados

corrplot(cor(Dados[,-c(1,9,10)]), method ='number')


## MODELO (com Idade, Sh, Soot, Dist, FK, PB e xG) ##
library(MASS)
library(car)
library(olsrr)
library(ggrepel)
library(dplyr)

library(gridExtra)
library(ggplot2)


## Modelo ##
modelo <- glm.nb(Gols ~ Idade + Sh + Soot + Dist +
                   FK + PT + xG, data = Dados, maxit = 34, link=log)
modelo

#### observação1: parâmetro init.theta: Valor inicial
opcional para o parâmetro phi.
#### Se omitido, um estimador de momento após um ajuste
```

```
inicial usando um Poisson GLM é usado.

#### Um processo de iteração alternada é usado. Para
determinado phi,
#### o GLM é ajustado usando o mesmo processo usado por glm().
#### Para médias fixas, o parâmetro phi é estimado
usando pontuações e iterações
#### de informações. Os dois são alternados até a
convergência de ambos.
#### (O número de alternâncias e o número de iterações
ao estimar phi são
#### controlados pelo parâmetro maxit de glm.control.)

## Observação2: até maxit = 33 o modelo apresenta aviso
de limite de alternância atingido
## Observação3: depois de maxit = 33 o parâmetro phi
fica constante (3.732626849)

anova(modelo) #TRV
# IC TRV
confint(modelo,level=0.95)

summary(modelo) #wald
# IC Wald
confint.default(modelo)

# seleção de variáveis: stepwise pelo critério AIC
stepAIC(modelo, direction = "both")

# seleção de variáveis: lasso
library(glmnet)
Data <- data.frame(Dados$Idade,Dados$Sh,Dados$Soot,
                   Dados$Dist,Dados$FK,Dados$PT,Dados$xG)
```

```
Data2 <- data.matrix(Data)
newfit <- glmnet(Data2,Gols, family = negative.binomial(theta = 3.732626849))
newfit

# Selecionar o melhor lambda
cv.newfit <- cv.glmnet(Data2,Dados$Gols, family =
negative.binomial(theta = 3.732626849))
plot(cv.newfit)
(best.lambda <- cv.newfit$lambda.min) ## 0.06945602

# Variáveis selecionadas
coef(newfit)[,38]

## Modelo stepwise ##
modelo2 <- glm.nb(Gols ~ Soot + Dist + FK  + PT + xG,
data = Dados, maxit = 34, link=log)
modelo2
modelo2$coefficients

anova(modelo2) #TRV

## Modelo lasso ##
modelo3 <- glm.nb(Gols ~ Soot + Dist + FK  + xG, data =
Dados, maxit = 34, link=log)
modelo3
modelo3$coefficients



anova(modelo3) #TRV

### ANALISE DE DIAGNÓSTICO PARA MODELO STEPWISE###
```

```
## Ponto de Alavanca e pontos influentes ##

## hii
Dados["obs"] <- c(seq(1,272))
da=data.frame(obs=c(1:272),hii=hatvalues(modelo2))

## Gráfico hii
(h <- ggplot(data = da,aes(x = obs, y = hii)) +
    geom_point() +
    geom_segment(aes(x = obs, xend = obs, y = 0, yend = hii)) +
    geom_hline(aes(yintercept =2*6/272), col = "orange",
               size = 0.8,linetype ="dashed") +labs(y = "hii",
    x = "Observações", colour = "Observação") +
    geom_text_repel(data = filter(da), color = "RED",
                    label = filter(da)$obs) +
    theme_light())

## Distância de cook
Di <- cooks.distance(modelo2)

## Gráfico de cook

(d <- ggplot(Dados ,aes(x = obs, y = Di)) +
    geom_point() +
    geom_segment(aes(x = obs, xend = obs, y = 0, yend = Di)) +
    labs(y = "Di", x = "Observações",colour = "Observação") +
    geom_text_repel(data = filter(Dados), color = "RED",
                    label = filter(Dados)$obs) + theme_light())


Dados2 <- Dados[-c(204, 232),]

###### Adequação da Função de Ligação ######
```

```
# funções de ligação: log

modelo4 <- glm.nb(Gols ~ Soot + Dist + FK  + PT + xG,
data = Dados2, maxit = 45, link= log)

# Resíduos
res <- residuals.glm(modelo4)
# Preditor linear
eta <- model.matrix(modelo4)%*%modelo4$coefficients
# Variável dependente ajustada
mu <- modelo4$fitted.values

## Gráfico de resíduos vs Preditor linear (Verifica se
função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta,res=round(res,3))
graf1 <- ggplot(df,aes(x=eta, y=res))+
          geom_point(size=1.5,shape=19, color = "black")+
          geom_hline(yintercept=0, colour="red", size=1)+
          theme_light()+
          labs(x=expression(paste(eta)), y = "Resíduos") +
          theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
          axis.text = element_text(size = 13), title =
          element_text(size = 15))

## Gráfico variavel dependente ajustada vs Preditor
linear (Verifica se função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta,mu=round(mu,3))
graf2 <- ggplot(df,aes(x=eta, y=mu))+
         geom_point(size=1.5,shape=19, color = "black")+
```

```
        theme_light()+
        labs(x=expression(paste(eta)), y = "Variável Dependente") +
        theme(plot.title = element_text(hjust = 0.5), axis.title =
        element_text(size = 13),
        axis.text = element_text(size = 13), title =
        element_text(size = 15))

library(gridExtra)
grid.arrange(graf1,graf2, ncol=2)

# Gráfico de dispersão com Soot

new_y <- log(Dados2$Gols)
a <- (ggplot(Dados2, aes(Dados2$Soot, new_y)) +
    geom_point(alpha = 0.6, color = "black", size = 1.5, shape = 19) +
    labs(x = "X", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19)) +
      theme_light())

# Gráfico de dispersão com Dist

new_y <- log(Dados2$Gols)
b <- (ggplot(Dados2, aes(Dados2$Dist, new_y)) +
    geom_point(alpha = 0.6, color = "black", size = 1.5, shape = 19) +
    labs(x = "X", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))+
      theme_light())
```

```
# Gráfico de dispersão com FK

new_y <- log(Dados2$Gols)
c <- (ggplot(Dados2, aes(Dados2$FK, new_y)) +
     geom_point(alpha = 0.6, color = "black", size = 1.5, shape = 19) +
     labs(x = "X", y = "Quantidade de gols marcados") +
     theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))+
       theme_light())

# Gráfico de dispersão com PT

new_y <- log(Dados2$Gols)
d <- (ggplot(Dados2, aes(Dados2$PT, new_y)) +
    geom_point(alpha = 0.6, color = "black", size = 1.5, shape = 19) +
    labs(x = "X", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
          axis.text = element_text(size = 13), title =
            element_text(size = 19))+
      theme_light())

# Gráfico de dispersão com xG

new_y <- log(Dados2$Gols)
e <- (ggplot(Dados2, aes(Dados2$xG, new_y)) +
    geom_point(alpha = 0.6, color = "black", size = 1.5, shape = 19) +
    labs(x = "X", y = "Quantidade de gols marcados") +
    theme(plot.title = element_text(hjust = 0.5), axis.title =
            element_text(size = 15),
```

```
          axis.text = element_text(size = 13), title =
            element_text(size = 19))+
      theme_light())

library(gridExtra)
grid.arrange(a,b,c,d,e, nrow = 2, ncol=3)


summary(Dados$Gols)
summary(Dados$Soot)
summary(Dados$Dist)
summary(Dados$FK)
summary(Dados$PT)
summary(Dados$xG)


## Identificando presença de multicolinearidade ##
library(car)
car::vif(modelo4)

## Envelope ##
library(hnp)
hnp(modelo4, main="Envelope", how.many.out = T)

### ANALISE DE DIAGNÓSTICO PARA MODELO LASSO ###

## Ponto de Alavanca e pontos influentes ##

## hii
Dados["obs"] <- c(seq(1,272))
da=data.frame(obs=c(1:272),hii=hatvalues(modelo3))

## Gráfico hii
```

```
(h <- ggplot(data = da,aes(x = obs, y = hii)) +
    geom_point() +
    geom_segment(aes(x = obs, xend = obs, y = 0, yend = hii)) +
    geom_hline(aes(yintercept =2*6/272), col = "orange",
               size = 0.8,linetype ="dashed") +labs(y = "hii",
    x = "Observações", colour = "Observação") +
    geom_text_repel(data = filter(da), color = "RED",
                    label = filter(da)$obs) +
    theme_light())

## Distância de cook
Di <- cooks.distance(modelo3)

## Gráfico de cook

(d <- ggplot(Dados ,aes(x = obs, y = Di)) +
    geom_point() +
    geom_segment(aes(x = obs, xend = obs, y = 0, yend = Di)) +
    labs(y = "Di", x = "Observações",colour = "Observação") +
    geom_text_repel(data = filter(Dados), color = "RED",
                    label = filter(Dados)$obs) + theme_light())

grid.arrange(h,d, ncol=2)


## Suposição de Independência e adequação da Função de Ligação ##

modelo5 <- glm.nb(Gols ~ Soot + Dist + FK  + xG, data =
Dados2, maxit = 34, link=log)

# Resíduos
res <- residuals.glm(modelo5)
# Valores preditos
```

```
pred <- log(fitted.values(modelo5))
# Preditor linear
eta=model.matrix(modelo5)%*%modelo5$coefficients
# Variável dependente ajustada
mu <- modelo5$fitted.values

## Gráfico de resíduos vs Preditor linear (Verifica se função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta,res=round(res,3))
m <- ggplot(df,aes(x=eta, y=res))+
  geom_point(size=1.5,shape=19, color = "black")+
  geom_hline(yintercept=0, colour="red", size=1)+
  theme_light()+
  labs(x=expression(paste(eta)), y = "Resíduos") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

## Gráfico variavel dependente ajustada vs Predito
linear (Verifica se função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta,mu=round(mu,3))
n <- ggplot(df,aes(x=eta, y=mu))+
  geom_point(size=1.5,shape=19, color = "black")+
  theme_light()+
  labs(x=expression(paste(eta)), y = "Variável Dependente") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

grid.arrange(m,n, ncol=2)
```

```
## Identificando presença de multicolinearidade ##
library(car)
car::vif(modelo5)
vif(modelo5)
## Envelope ##
library(hnp)
hnp(modelo5, main="Envelope",resid.type = "deviance",how.many.out = T)

# funções de ligação: identidade

modelo6 <- glm.nb(Gols ~ Soot + Dist + FK + xG, data =
Dados2, maxit = 45, link= identity)
## Não funciona



# funções de ligação: sqrt

modelo7 <- glm.nb(Gols ~ Soot + Dist + FK + PT + xG,
data = Dados2,  maxit = 45, link= sqrt)

# Resíduos
res2 <- residuals.glm(modelo7)
# Preditor linear
eta2 <- model.matrix(modelo7)%*%modelo7$coefficients
# Variável dependente ajustada
mu2 <- modelo7$fitted.values

## Gráfico de resíduos vs Preditor linear (Verifica se
função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta2,res2=round(res2,3))
```

```
graf3 <- ggplot(df,aes(x=eta2, y=res2))+
  geom_point(size=1.5,shape=19, color = "black")+
  geom_hline(yintercept=0, colour="red", size=1)+
  theme_light()+
  labs(x=expression(paste(eta)), y = "Resíduos") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

## Gráfico variavel dependente ajustada vs Preditor
linear (Verifica se função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta2,mu2=round(mu2,3))
graf4 <- ggplot(df,aes(x=eta2, y=mu2))+
  geom_point(size=1.5,shape=19, color = "black")+
  theme_light()+
  labs(x=expression(paste(eta)), y = "Variável Dependente") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

library(gridExtra)
grid.arrange(graf3,graf4, ncol=2)

modelo7 <- glm.nb(Gols ~ Soot + Dist + FK + PT + xG,
data = Dados2,  maxit = 45, link= sqrt)
## Identificando presença de multicolinearidade ##
library(car)
car::vif(modelo7)

library(pscl)
```

```
m1ZBN <- zeroinfl(Gols ~ Soot + FK + PT + xG, data = Dados2, dist = "negbin")
## Identificando presença de multicolinearidade ##
library(faraway)
car::vif(m1ZBN)
hnp(m1ZBN, main="Envelope",how.many.out = T)


## Envelope ##
library(hnp)
hnp(modelo7, main="Envelope",resid.type = "deviance",how.many.out = T)

## Ajustando modelo com distribuição Poisson e função de ligação logaritmica ##

modelo_poisson2 <- glm(Gols ~ Soot + Dist + FK  + PT +
xG, data = Dados2, family = poisson(link = "log"))

# Resíduos
res_poisson2 <- residuals.glm(modelo_poisson2)
# Preditor linear
eta_poisson2 <- model.matrix(modelo_poisson2)%*%modelo_poisson2$coefficients
# Variável dependente ajustada
mu_poisson2 <- modelo_poisson2$fitted.values


## Gráfico de resíduos vs Preditor linear (Verifica se
função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta,res=round(res_poisson2,3))
y <- ggplot(df,aes(x=eta_poisson2, y=res_poisson2))+
  geom_point(size=1.5,shape=19, color = "black")+
  geom_hline(yintercept=0, colour="red", size=1)+
  theme_light()+
  labs(x=expression(paste(eta_poisson2)), y = "Resíduos") +
```

```
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

## Gráfico variavel dependente ajustada vs Preditor
linear (Verifica se função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta_poisson2,mu_poisson2=round(mu_poisson2,3))
z <- ggplot(df,aes(x=eta_poisson2, y=mu_poisson2))+
  geom_point(size=1.5,shape=19, color = "black")+
  theme_light()+
  labs(x=expression(paste(eta_poisson2)), y = "Variável Dependente") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

grid.arrange(y,z, ncol=2)

## Envelope ##
library(hnp)
hnp(modelo_poisson2, main="Envelope",how.many.out = T)

## Ajustando modelo com distribuição Poisson e função de ligação raiz quadrada ##

modelo_poisson <- glm(Gols ~ Soot + Dist + FK  + PT +
xG, data = Dados2, family = poisson(link = "sqrt"))

# Resíduos
res_poisson <- residuals.glm(modelo_poisson)
# Preditor linear
eta_poisson <- model.matrix(modelo_poisson)%*%modelo_poisson$coefficients
```

```
# Variável dependente ajustada
mu_poisson <- modelo_poisson$fitted.values


## Gráfico de resíduos vs Preditor linear (Verifica se
função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta,res=round(res_poisson,3))
y <- ggplot(df,aes(x=eta_poisson, y=res_poisson))+
  geom_point(size=1.5,shape=19, color = "black")+
  geom_hline(yintercept=0, colour="red", size=1)+
  theme_light()+
  labs(x=expression(paste(eta_poisson)), y = "Resíduos") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

## Gráfico variavel dependente ajustada vs Preditor
linear (Verifica se função de ligação é adequada)
library(gghighlight)
df <- data.frame(eta_poisson,mu_poisson=round(mu_poisson,3))
z <- ggplot(df,aes(x=eta_poisson, y=mu_poisson))+
  geom_point(size=1.5,shape=19, color = "black")+
  theme_light()+
  labs(x=expression(paste(eta_poisson)), y = "Variável Dependente") +
  theme(plot.title = element_text(hjust = 0.5), axis.title =
          element_text(size = 13),
        axis.text = element_text(size = 13), title =
          element_text(size = 15))

grid.arrange(y,z, ncol=2)
```

```
## Envelope ##
library(hnp)
hnp(modelo_poisson, main="Envelope",how.many.out = T)

## MODELOS FINAIS ##
# Modelo binomial negativa final
modelo_bn <- glm.nb(Gols ~ Soot + Dist + FK + xG, data
= Dados2,  maxit = 45, link= sqrt)
summary(modelo_bn)

# Modelo poisson final
modelo_poisson3 <- glm(Gols ~ Soot + Dist + FK + xG,
data = Dados2, family = poisson(link = "sqrt"))
summary(modelo_poisson3)
```