Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

REGRESSÃO LOGÍSTICA NA MODELAGEM DA PROBABILIDADE DE VITÓRIA EM JOGOS DE FUTEBOL AMERICANO

Pablo Henrique de Freitas

Uberlândia-MG
2019

Pablo Henrique de Freitas

REGRESSÃO LOGÍSTICA NA MODELAGEM DA PROBABILIDADE DE VITÓRIA EM JOGOS DE FUTEBOL AMERICANO

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Profª Dra. Maria Imaculada de Sousa Silva

Uberlândia-MG

2019

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Profª Dra. Maria Imaculada de Sousa Silva

_____

Prof. Dr. José Waldemar Silva

_____

Prof. Dr. Rogério de Melo Costa Pinto

Uberlândia-MG

2019

Agradecimentos

Agradeço, primeiramente a Deus, que com sua maravilhosa Graça, concedeu-me tudo o que foi necessário para passar por esses 4 anos e meio de graduação com êxito. Toda minha gratidão a Ele!

À minha mãe Rosilene e ao meu pai Paulo, por todo apoio, amor, educação, ensinamentos e sustento. Sem todos os esforços e sacrifícios feitos por eles, nada disso seria possível.

Ao meu professor de matemática do Ensino Fundamental e Médio, Alessandro Moreira, que me fez gostar de matemática e me apresentou a Estatística. Graças às suas dicas, orientações e apoio, pude descobrir e me apaixonar pela minha amada profissão.

Aos professores Rogério de Melo, José Waldemar e Maria Imaculada, pelas orientações em trabalhos, projetos de pesquisa e trabalho de conclusão do curso, realizados durante a graduação. Tudo isso foi muito importante para o desenvolvimento das minhas habilidades analíticas e conhecimento prático.

A todos os meus colegas e amigos que de alguma forma contribuiram comigo durante o curso.

Resumo

O presente trabalho tem por objetivo identificar por meio da análise de regressão logística quais fatores podem influenciar no resultado final de uma partida de futebol americano da NFL (National Football League). Para tanto, foram coletados dados referentes a 413 partidas da Liga das temporadas de 2014 a 2015 no site oficial da NFL, e fez-se o uso de um modelo logístico, tratando como evento resposta o resultado das equipes mandantes (vitória ou derrota) nos jogos. Foram analisadas 38 variáveis que são possíveis fatores e características das equipes, que podem influenciar no resultado de uma partida. Para avaliar e medir a qualidade do ajuste do modelo, foram realizados alguns testes de diagnóstico como o Teste de Hosmer-Lemeshow, Deviance e Pearson. O modelo final se mostrou bem ajustado aos dados, com uma boa capacidade discriminatória, com um valor de AUC (*Area Under the ROC Curve*) de 0,718 acertando cerca de 60% das previsões realizadas. Observou-se também que dentre as variáveis estudadas apenas a média de pontos marcados por partida da equipe mandante, a média de jardas de passe conquistadas por partida pela equipe mandante, a porcentagem de conversões em terceiras descidas da defesa da equipe da casa, o aproveitamento de vitórias na temporada da equipe visitante e o resultado do último jogo dos visitantes apresentaram influencia significativa na probabilidade de uma equipe vencer uma partida da NFL.

Palavras-chave: curva roc, futebol americano, modelo preditivo, Regressão logistica, testes de diagnóstico.

Abstract

The present work aims to identify through logistic regression analysis which factors can influence the final result of an NFL football game (National Football League). In order to do so, we collected data on 413 league matches from the 2014 to 2015 seasons on the official NFL website, and used a logistic model, treating as an answer event the result of the main teams (victory or defeat) in the games. We analyzed 38 variables that are possible factors and characteristics of the teams, which can influence the outcome of a match.

To evaluate and measure the quality of fit of the model, some diagnostic tests such as the Hosmer-Lemeshow, Deviance and Pearson Test were performed. The final model was well-adjusted to the data, with a good discriminatory capacity, with an AUC value of 0.718 matched around 60% of the predictions made. It was also observed that among the studied variables only the average points scored by the home team, the average of yards of pass earned by the home team, the % of conversions in third descents of the defense of the home team, team victories in the visiting team in the season and the outcome of the visitors last game had a significant influence on the likelihood of a team winning an NFL game.

Keywords: ROC curve, american football, predictive models, logistic regression,diagnostic tests .

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Assim como o futebol, esporte de maior popularidade no Brasil, o futebol americano vem ganhando muita popularidade entre os brasileiros nos últimos anos. E, também como no futebol, análises e previsões acerca desse esporte são ainda um universo obscuro para muitos especialistas. Por ser um esporte complexo, com vários jogadores em campo e vários eventos aleatórios que podem influenciar nos resultados das partidas e das jogadas, previsões nesse cenário são feitas apenas no campo subjetivo, baseado-se apenas no conhecimento técnico e "achismo"dos opinadores [14].

Com surgimento no século XIX, o Futebol Americano atualmente é o esporte mais popular dos Estados Unidos. Cada equipe basicamente possui três times, um de ataque, um de defesa e um de especiais, que se revezam em campo de acordo com o contexto do jogo, sempre com 11 jogadores de cada lado. O objetivo principal do jogo é marcar *Touchdowns*, o que concede à equipe 7 pontos no placar. Os jogadores de ataque têm 4 tentativas para avançar 10 jardas no campo, e assim renovar suas tentativas para chegar a *End Zone* do campo do adversário, marcando um *Touchdown*. Para fazer isso, a equipe pode tanto passar a bola como correr com ela, o que torna o esporte um dos mais estratégicos e complexos do mundo.

Nos Estados Unidos, todos os esportes ganham atenção especial no que se diz respeito a números e coleta de dados. Em todos os jogos, são geradas milhares de informações que auxiliam treinadores, comentaristas esportivos e analistas de desempenho em suas atividades.

Nesse contexto, fica claro o interesse de jogadores e treinadores em saber quais das centenas de fatores que podem influenciar no resultado de uma partida realmente têm influência significativa. Sabendo quais características de fato importam, treinadores podem trabalhar junto à sua comissão técnica ações e treinamentos focados em tais fundamentos de jogo de seus jogadores.

Também é interesse de apostadores esportivos profissionais prever probabilisticamente, com a maior assertividade possível, resultados de eventos esportivos [7]. As casas de apostas desportivas contam com um aparato tecnológico, estatístico e grande quantidade de dados, permitindo prever milhares de eventos esportivos com a maior precisão possível, fato esse que a coloca em vantagem em relação aos apostadores. Essa vantagem se dá pelo fato de ela desvalorizar a aposta para garantir seu lucro, fazendo com que o apostador mesmo vencendo a maioria de suas apostas tenha prejuízo a longo prazo.

Técnicas estatísticas de regressão e estimação utilizando modelos probabilísticos são as ferramentas mais utilizadas para se fazer previsões acerca de eventos esportivos. Olhando para

a literatura, pode-se perceber que modelos estatísticos são de grande valia para se realizar inferências acerca de alguns esportes, pois há uma quantidade razoável de trabalhos publicados em que vários autores utilizam modelos de probabilidade para prever eventos esportivos, principalmente o futebol.

O maior evento esportivo do mundo, a Copa do Mundo de Futebol FIFA, obviamente foi alvo de vários estudos de previsão de seus resultados, por parte de alguns autores, como por exemplo os estudos de Dyte e Clarke (2000) e Suzuki *et al.* (2010). Dyte e Clarke (2010) utilizaram um modelo de regressão de Poisson, considerando informações relacionadas à força de ataque e defesa das equipes, além da classificação das equipes no ranking da FIFA (*Federation Internationale of Football Association*) [5].

Com uma abordagem estatística diferente, Suzuki *et al.* (2010) [2] propôs um modelo de regressão Bayesiano para prever as probabilidades dos resultados das partidas utilizando as opiniões de especialistas e rankings da FIFA como informação prévia.

Lee (1997) [11] desenvolveu um modelo de regressão de Poisson para prever as probabilidades de vitória, empate e derrota das equipes em partidas da *Premier League* (Liga de futebol nacional da Inglaterra), utilizando-se como informação das equipes dados como média de gols marcados e sofridos. O autor realizou também, previsões acerca da pontuação final das equipes ao fim do campeonato.

Karlis e Ntzoufras (2003) [10] foi um pouco além, e realizou previsões de partidas de futebol utilizando modelos de Poisson bivariados. As informações utilizadas pelo autor para a estimação dos parâmetro do modelo também foram relacionadas com o poder de ataque e defesa das equipes envolvidas na partida, sendo utilizados dados dos times da liga nacional de futebol da Itália (Serie A) da temporada de 1991-1992.

Modelos estatísticos também são utilizados para realizar previsões em esportes não tão populares, como por exemplo, o Polo aquático. Karlis e Ntzoufras (2003) [10] utilizou modelos de Poisson para realizar previsões e simulações de jogos da European National Cup de 1999.

Nesse cenário, modelos estatísticos surgem como uma ferramenta poderosa no auxilio à previsões e explanações acerca do futebol americano. Com um lago de dados à disposição de forma livre em diversos sites americanos, estudos probabilísticos nessa área são uma maneira confiável para se fazer predições. Prever o resultado de uma partida da NFL pode não ser algo simples e fácil dado toda a complexidade que envolve o jogo, por isso a atribuição de modelos preditivos se torna uma opção viável para relacionar covariáveis que traduzem melhor as chances de uma equipe sair vencedora de um confronto.

Como observado em muitos dos trabalhos de previsões esportivas, os modelos de Poisson geralmente são muito utilizados, porém, devido à quantidade de fatores e forma de pontuação do futebol americano, pode ser mais simples e efetivo trabalhar com um modelo de regressão logística.

De acordo com Hosmer e Lemeshow (2000) [8], quando se tem uma variável resposta categórica ou binária é possível ajustar um modelo de regressão logístico. Sendo assim, o objetivo geral do presente trabalho é ajustar um modelo de regressão logística, onde seja possível identi-

ficar quais covariáveis possuem efeito significativo no resultado final em uma partida de futebol americano, e prever probabilisticamente qual é o time com mais chances de vencer determinado confronto.

2. Fundamentação Teórica

2.1 Modelo de regressão logística

Nos modelos de regressão em geral, estuda-se a relação de uma variável resposta Y com outras variáveis independentes $(X_1, X_2, ..., X_p)$, em que a média de Y é definida por $E(Y|X)$, sendo X o vetor que contém as variáveis preditoras.

Os dois tipos de modelos mais conhecidos e utilizados são os de regressão Linear e Logística. Enquanto nos modelos lineares a variável dependente é contínua e a estimação dos parâmetros é feita pelo método dos mínimos quadrados, nos modelos logísticos a variável resposta é sempre categórica e a estimação é feita pelo método da máxima verossimilhança. Além disso, os modelos logísticos não necessitam de algumas pressuposições que os modelos lineares exigem, como por exemplo, a normalidade dos erros e a homogeneidade da variância [4].

Os modelos de regressão logística são comumente utilizados em problemas de classificação, em situações que a variável dependente é de natureza dicotômica ou binária, onde as variáveis independente podem ser categóricas ou não. Através dessa técnica, é possível também estimar a probabilidade associada à ocorrência de um determinado evento em função do conjunto de variáveis explicativas.

Na regressão logística binária, a variável resposta Y tem distribuição Bernoulli, em que $\pi(x) = P(Sucesso) = P(Y = 1|X)$. A variável Y só assume dois valores, 0 ou 1, sendo 1 a ocorrência do evento de interesse com probabilidade $\pi(x)$ e 0 a ocorrência do não evento com probabilidade $[1 - \pi(x)]$. Nesse modelo, as variáveis independentes X_k, sendo $k = 1, 2, ..., p$, podem ser categóricas ou contínuas[8].

O modelo de regressão logística é definido pela expressão:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}} \tag{2.1}$$

onde $\pi(x) = E(Y|X)$ e p é o número de variáveis independentes consideradas no modelo ajustado, e o preditor linear é $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.

Os coeficientes $\beta_0, \beta_1, ..., \beta p$ são estimados pelo método da máxima verossimilhança a partir do conjunto de dados, sendo que os valores de obtidos de $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta} p$ maximizam o logaritmo da função de verossimilhança.

Pode-se observar na Figura 2.1 que a curva logística para a relação de uma variável Y com uma covariável x_k tem um comportamento probabilístico no formato de "S", fato esse que é algo

característico dos modelos de regressão logística com coeficiente positivo (Hosmer e Lemeshow, 2000). Há também a curva logística com formato de "S"invertido, sendo o oposto da Figura 2.1, que ocorre em casos onde o coeficiente β é negativo.

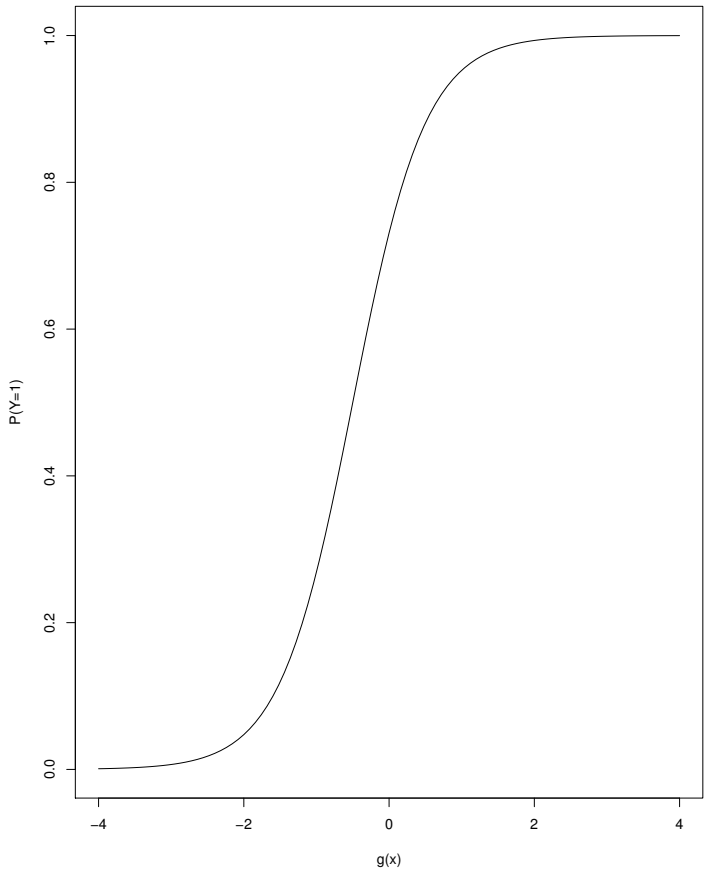


Figura 2.1: Função característica do modelo de regressão logística binária.

A interpretação dos parâmetros de um modelo de regressão logística é feita comparando a probabilidade de sucesso com a probabilidade de fracasso, utilizando a função denominada *odds ratio* ou razão de chances (OR). Calcula-se essa função a partir da função *odds*, onde:

$$g(x) = \frac{\pi(x)}{[1-\pi(x)]} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_i} \tag{2.2}$$

Desta forma, tomando dois valores distintos de uma variável explicativa x_k qualquer, $x_{k(j+1)}$ e x_{kj}, temos:

$$OR = \frac{g(x_{k(j+1)})}{g(x_{kj})} \tag{2.3}$$

E, aplicando o logaritmo, obtemos:

$$ln(OR) = ln\left[\frac{g(x_{k(j+1)})}{g(x_{kj})}\right] = ln\left[g(x_{k(j+1)})\right] - ln\left[g(x_{kj})\right] = \beta_k(x_{k(j+1)} - x_{kj}) \tag{2.4}$$

Assim, considerando os valores das variáveis explicativas como $x_{j+1} - x_j = 1 und.$, temos

que:

$$ln(OR) = ln(e^{\beta_k}) = \beta_k \tag{2.5}$$

Desta maneira, tomando como referência os valores de uma variável explicatória x_i, podemos obter a chance de sucesso de um grupo em relação a outro utilizando a função *odds ratio*, em que cada um desses dois grupos possuem valores diferentes da mesma variável x_i, como por exemplo, homens comparados as mulheres, doentes comparados a saudáveis, entre outros.

$$OR(\beta_k) = e^{\beta_k} \tag{2.6}$$

Assim, quando $\beta_k > 0$ então $OR > 1$ e, consequentemente, as chances de sucesso dos indivíduos x_{j+1} são maiores que as dos indivíduos x_j. Por outro lado, quando $\beta_k < 0$ então $OR < 1$, e assim as chances de sucesso dos indivíduos x_{j+1} são menores que as dos indivíduos x_j.

2.2 Estimação dos parâmetros do modelo

A estimação dos parâmetros β_i do modelo de regressão logística é feita pelo método da máxima verossimilhança. Os valores de $\hat{\beta}_i$ que maximizam o logaritmo da função de verossimilhança são os estimadores de máxima verossimilhança de β_i. Considerando uma amostra independente (x_i, y_i) de tamanho n, em que x_i é o valor da variável independente da i-ésima observação e y_i é o valor da variável resposta dicotômica, Y_i $Ber(\pi_i)$, isto é, Y tem distribuição de probabilidade Bernoulli. A distribuição de probabilidade de Y_i é dada por:

$$f(y_i, \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{2.7}$$

Dado que as observações Y_i são independentes, a função de verossimilhança é dada pela seguinte expressão:

$$L(\beta) = \prod_{i=1}^{n} f(y_i, \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{2.8}$$

onde $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ é o vetor dos parâmetros do modelo, e y_i é 0 e 1.

Deseja-se determinar os estimadores $\hat{\beta}_i$ que maximiza $L(\beta)$ por meio da máxima verossimilhança. Portanto, aplicando o logaritmo na expressão (2.8) tem-se:

$$l(\beta) = ln(L(\beta)) = ln\left(\prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}\right) = \sum_{i=1}^{n}\left[y_i ln\left(\frac{\pi_i}{1 - \pi_i}\right) + ln(1 - \pi_i)\right] \tag{2.9}$$

Substituindo π_i e $1-\pi_i$ da equação (2.9) por suas respectivas expressões, tem-se:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i ln\left(\beta_0 + \beta_1 x_i + ... + \beta_p x_i\right) + ln\left(\frac{1}{1 + exp(\beta_0 + \beta_1 x_i + ... + \beta_p x_i)}\right)\right] \tag{2.10}$$

$$l(\beta) = \sum_{i=1}^{n} \left[y_i ln\left(\beta_0 + \beta_1 x_i + ... + \beta_p x_i\right) - ln\left(1 + exp(\beta_0 + \beta_1 x_i + ... + \beta_p x_i)\right)\right] \tag{2.11}$$

Derivando a expressão (2.11) em relação a cada parâmetro do modelo, serão obtidas $p+1$ equações de máxima verossimilhança para os $p+1$ parâmetros que deseja-se estimar. Essas equações podem ser expressas por:

$$\sum_{i=1}^{n} y_i ln\left[\pi(x_i)\right] = 0 \tag{2.12}$$

e

$$\sum_{i=1}^{n} x_{ij} ln\left[y_i - \pi(x_i)\right] = 0 \tag{2.13}$$

As equações de verossimilhança para a regressão logística, são não-lineares em β, logo, necessita-se de métodos especiais para solucioná-las. Tais métodos são iterativos e requerem auxílio computacional disponíveis na maioria dos softwares estatísticos, sendo o mais comum deles o método de estimação de Newton-Raphson.

2.3 TESTES DE AJUSTE E SELEÇÃO DO MODELO

Assim como todos os modelos estatísticos, é necessários realizar testes para verificar a qualidade de ajuste do modelo estimado, a detecção de pontos discrepantes (*outliers*) e garantir a confiabilidade do modelo.

Na regressão logística, pode-se avaliar a quão bem o modelo estimado se ajusta aos dados observados através de gráficos, porém, existem medidas como a *deviance*, a estatística qui-quadrado de Pearson, o teste de Hosmer e Lemeshow e o teste de Wald, que ajudam a verificar a qualidade de ajuste do modelo.

2.3.1 TESTE DE WALD

No processo de ajuste de um modelo de regressão logística, o Teste de Wald é utilizado para testar a significância de cada um dos parâmetros no modelo. Este teste considera a estimativa de máxima verossimilhança do parâmetro β_k com a estimativa do seu erro padrão dado por $\hat{EP}(\hat{\beta}_k)$. Sob a hipótese de que β_k é igual a zero, W segue distribuição Normal padrão.

A estatística de teste utilizada no Teste de Wald é dada pela expressão:

$$W = \frac{\beta_k}{\hat{E}P(\hat{\beta}_k)} \tag{2.14}$$

2.3.2 CRITÉRIO DE INFORMAÇÃO DE AKAIKE (AIC)

O critério de informação de Akaike (AIC) é um método de seleção de modelos, desenvolvido por meio dos estimadores de verossimilhança (EMV). Tal critério é utilizado para decidir qual o modelo mais adequado quando se compara modelos aninhados.

A decisão do melhor modelo de acordo com o critério é relativamente simples, onde o melhor modelo ajustado é o que possui o menor valor de AIC. A estatística AIC é definida como:

$$AIC = -2l(\beta) + 2(p+1) \tag{2.15}$$

onde $l(\beta)$ é o logarítmo da função de verossimilhança do modelo e p é o número de parâmetros estimados.

2.3.3 TESTE DE DEVIANCE

Na regressão logística, a expressão equivalente à soma dos quadrados dos resíduos da regressão linear é denominada $Deviance(D)$, onde o resíduo d é definido pela expressão:

$$d(y_j, \hat{\pi}_i) = sinal\left[2\left[y_i ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i) ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right]\right]^{1/2} \tag{2.16}$$

A estatística D, sob hipótese H_0 de que o modelo está bem ajustado, tem distribuição assintótica Qui-quadrado χ^2 com $n-(p+1)$ graus de liberdade, onde p é o número de covariáveis no modelo e o sinal é da equação é dado por $(y_i - m\hat{\pi})$. A estatística de teste é dada pela expressão:

$$D = \sum_{i=1}^{n} d(y_i, \hat{\pi}_i)^2 \tag{2.17}$$

O p-valor para o teste de Deviance tende a ser menor para os dados que estão no formato de Resposta/Frequência binária em relação aos dados no formato de Evento/Ensaio. Para os dados em formato de Resposta/Frequência binária, os resultados do teste Hosmer-Lemeshow são mais confiáveis [12].

2.3.4 TESTE DE PEARSON

Baseada nos resíduos de Pearson, a estatística de teste é obtida pela seguinte expressão:

$$\chi^2 = \sum_{j=i}^{n} r(y_i, \hat{\pi}_i)^2 \tag{2.18}$$

onde r_j é o resíduo de Pearson para o *r-ésimo* elemento, dado pela expressão:

$$r(y_i, \hat{\pi}_i) = r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}} \tag{2.19}$$

Sob a hipótese nula de que o modelo ajustado é adequado, a estatística de teste tem distribuição Qui-Quadrado com $n - (p + 1)$ g.l, em que p é o número de covariáveis no modelo.

A aproximação para a distribuição Qui-quadrado que utiliza-se no teste de Pearson é imprecisa quando o número esperado de eventos por linha nos dados é pequeno. Portanto, o teste de qualidade de ajuste de Pearson é impreciso quando os dados estão em formato de Resposta/Frequência binária.

2.3.5 TESTE DE HOSMER-LEMESHOW

O teste de Hosmer-Lemeshow é um dos mais utilizados em regressão logística com o objetivo de atestar a qualidade de ajuste do modelo, ou seja, o teste comprova se o modelo obtido explica adequadamente os dados observados da variável resposta. O teste é baseado na divisão dos dados em g grupos (geralmente $g = 10$) de acordo com as probabilidades previstas, como por exemplo, em uma subdivisão de 10, tem-se grupos com probabilidades entre 0 e 0,1, entre 0,1 e 0,2, e assim sucessivamente até o último grupo que tem as probabilidades previstas entre 0,9 e 1.

Sob a hipótese nula de que o modelo está bem ajustado, a estatística C segue distribuição χ^2 com t graus de liberdade. A estatística de teste é calculada pela expressão:

$$C = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{e_k\left(1 - \frac{e_k}{n_k}\right)} \tag{2.20}$$

onde o_k é o número de casos observados entre os *k-ésimos* decis e e_k é o número esperado de casos no *k-ésimo* decil.

O teste de Hosmer-Lemeshow não depende do número de ensaios por linha nos dados como os outros testes de qualidade de ajuste. Assim, quando os dados têm poucos ensaios por linha, o teste de Hosmer-Lemeshow é um indicador mais confiável de quão bem o modelo ajusta os dados.

2.4 ANÁLISE DE RESÍDUOS E DIAGNÓSTICO DO MODELO

No processo de desenvolvimento de um modelo preditivo, pode ocorrer que após a escolha de um modelo e subsequente ajuste a um conjunto de dados o resultado obtido seja insatisfatório. Tal fato pode ocorrer em função de algum desvio sistemático entre os valores observados e ajustados, ou porque um ou mais valores observados são discrepantes em relação aso demais.

Com isso, a detecção de uma escolha errada da função de ligação pode acontecer por ela estar realmente errada, ou por uma ou mais covariáveis estarem na escala errada ou devido a

presença de pontos discrepantes no conjunto de dados. Tudo isso faz com que a verificação da adequação do modelo para um determinado conjunto de dados seja um processo difícil.

As técnicas para verificação do ajuste de um modelo podem ser formais ou informais, sendo que as informais se baseiam em exames visuais de gráficos para detecção de pontos aberrantes ou padrões.

Nesta seção são apresentados algumas técnicas e gráficos de diagnóstico e análise de resíduos utilizados para verificar a qualidade de ajuste do modelo aos dados.

2.4.1 LEVERAGE

Em um modelo de regressão linear normal, uma medida de alavancagem é dada pelos elementos da diagonal da matriz H, conhecida como matriz *Hat* ou matriz de projeção, dada pela expressão:

$$H = X(X'X)^{-1}X' \tag{2.21}$$

Os resíduos $\hat{\epsilon} = y - \hat{y}$, podem ser expressos em função da matriz H, pois $\hat{\epsilon} = (I - H)y$, onde I é a matriz identidade. Pregibon (1981) [15] realizou uma aproximação linear para os valores ajustados usando a regressão de mínimos quadrados ponderados como modelo, definindo assim a matriz H para a regressão logística por:

$$H = V^{\frac{1}{2}}X(X'VX)^{-1}X'V^{\frac{1}{2}} \tag{2.22}$$

onde V é uma matriz diagonal JxJ composta pelos elementos $v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)]$.

Desta forma, os pontos de alavanca para o modelo de regressão logística são denotados por h_j para a j-ésima diagonal de H, conforme expressão abaixo:

$$h_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)](I, x_j)(X'VX)^{-1}(I, x_j)' \tag{2.23}$$

Pelo fato da matriz H ser simétrica e idempotente, tem-se que $0 \leq h_j \leq 1$.

Os pontos de alavanca (h_j) medem o quão distante a observação x_j está das demais observações do conjunto de dados. O elemento h_j só depende dos valores das variáveis explicativas, ou seja, da matriz X, não envolvendo as observações y. Portanto, se h_j é grande, os valores das variáveis explicativas da j-ésima observação são discrepantes, ou seja, estão distantes dos valores médios das variáveis explicativas.

Tais pontos podem exercer um papel importante na estimação dos parâmetros de um modelo, sendo que sua exclusão pode implicar mudanças dentro de uma análise estatística. Alguns autores, como Montgomery, sugerem que pontos com h_j maior ou igual a $2(p+1)/n$ devem ser investigados, sendo p o número de parâmetros do modelo estimado [4].

2.4.2 DFBETAS

A medida $DFBETA_j$, indica para cada coeficiente βj relacionada a um preditor X_j o quanto ele se modifica quando a observação j é excluída. A estatística $DFBETA_j$ é definida pela expressão:

$$\hat{\beta} - \hat{\beta_{(j)}} = (X'X)^{-1}x'_j(1-h_j)^{-1}\hat{\epsilon_j} \tag{2.24}$$

O sinal do DFBETA indica se a inclusão de uma observação leva a um aumento ou decréscimo do coeficiente estimado de β_j, e o seu valor absoluto mostra o tamanho dessa diferença em relação ao desvio padrão estimado. De acordo com Neter et al (1996) [9], é necessário investigar observações que apresentarem valores absolutos de DFBETAS maiores que 1 para amostras pequenas, e no caso de amostras grande os valores absolutos de DFBETAS maiores que $2/\sqrt{n}$.

2.5 Desempenho de predição do modelo

Atualmente na ciência de dados, algumas métricas são utilizadas para se avaliar modelos de classificação, como o desenvolvido no presente trabalho. Para analisar se o modelo estimado previu bem os dados, se ele pode prever bem a classe de interesse, entre outras questões, pode-se utilizar ferramentas como a Matriz de confusão, a curva ROC, e algumas métricas como o Recall, Precisão e Acurácia, que serão apresentadas nesta seção.

2.5.1 Curva ROC e AUC

Quanto temos um modelo estatístico com a variável resposta binária, é necessário escolher uma regra de predição ($\hat{Y} = 0$ ou 1), dado que $\hat{\pi}$ esta entre 0 e 1. É intuitivo pensar que quanto maior o valor de $\hat{\pi}$, $\hat{Y_i} = 1$, e se $\hat{\pi}$ for mais próximo do 0, $\hat{Y_i} = 0$. Para determinar o ponto para o qual os valores abaixo dele o indivíduo é classificado como não evento ($\hat{Y_i} = 0$) e para os valores acima dele o indivíduo é classificado como evento ($\hat{Y_i} = 1$), é obtido um valor chamado de ponto de corte [1].

Um meio muito utilizado para determinar o ponto de corte, é a Curva ROC (Receiver Operating Characteristic Curve), que plota a sensibilidade do modelo contra 1-especificidade, para todos os possíveis pontos de corte entre 0 e 1.

A sensibilidade, também chamada de verdadeiros positivos (true positive - TP), avalia a capacidade de o modelo classificar um indivíduo como evento dado que ele realmente é evento, ou seja, $P(\hat{Y} = 1|Y = 1)$. A especificidade, também conhecida como verdadeiros negativos (true negative - TN), mede a capacidade de o modelo predizer um indivíduo como não evento sendo que ele realmente é não evento, isto é, $P(\hat{Y} = 0|Y = 0)$.

Escolhido o ponto de corte, que é o ponto da curva mais próximo da extremidade superior esquerda do gráfico (0,1), para facilitar a análise da curva ROC e medir o poder de discriminação do modelo, isto é, discriminar os indivíduos eventos dos não eventos, a AUC (*Area Under the*

ROC Curve), isto é, a área abaixo da curva ROC, pode ser utilizada. Esta métrica varia entre 0 e 1, sendo que quanto maior o valor de AUC, melhor a capacidade de classificação do modelo.

Na Figura 2.2, pode-se observar um exemplo com a forma de uma curva ROC, onde a reta diagonal que cruza os pontos (0,0) e (1,1) é a linha referência para um modelo ruim, isto é, um modelo com discriminação aleatória.

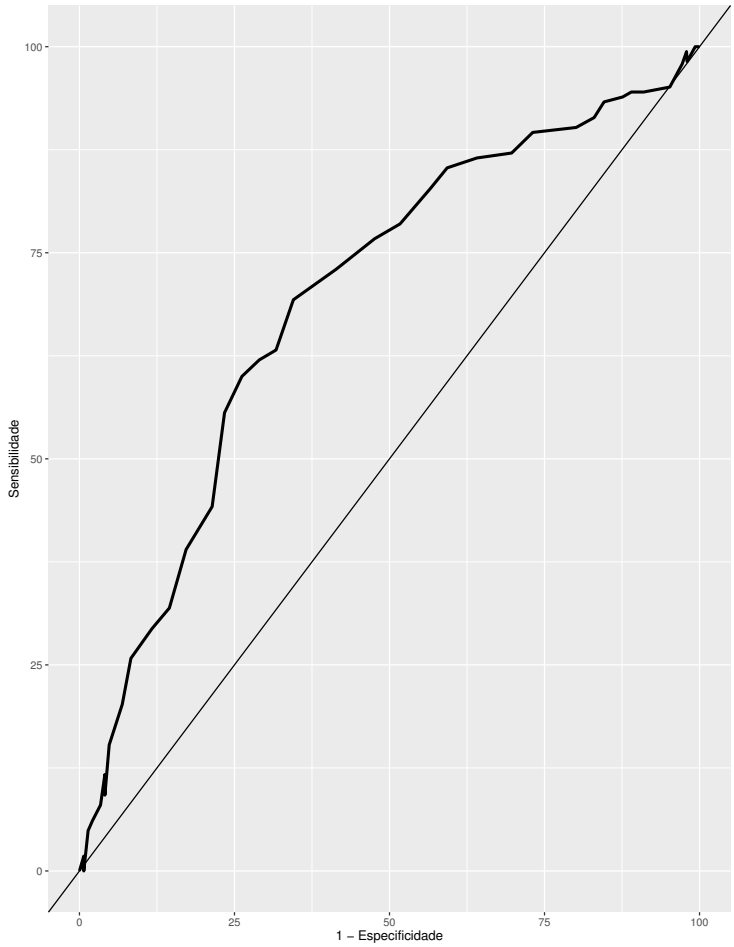


Figura 2.2: Um exemplo de curva ROC.

2.5.2 Matriz de confusão e métricas de desempenho

A maioria das métricas de desempenho de predição de um modelo estatístico são derivadas da matriz de confusão, que é uma tabela que mostra as frequências de classificação para cada classe predita e observada do modelo [1]. A matriz de confusão é composta por quatro medidas, que são elas:

- Verdadeiro positivo (true positive - TP): indivíduo classificado corretamente pelo modelo como evento dado que é realmente um evento.
- Falso positivo (false positive - FP): indivíduo classificado incorretamente pelo modelo como evento dado que é um não evento.

- Verdadeiro negativo (true negative - TN): indivíduo classificado corretamente pelo modelo como não evento dado que é realmente um não evento.
- Falso negativo (false negative - FN): indivíduo classificado incorretamente pelo modelo como não evento dado que é um evento.

Dado essas informações, a matriz de confusão é construída conforme observado na Tabela 2.1.

Tabela 2.1: Matriz de confusão

	Valores observados	
Valores Preditos	Evento	Não evento
Evento	TP	FP
Não evento	FN	TN

A partir da matriz de confusão, é possível calcular algumas medidas que são úteis para avaliar o desempenho preditivo do modelo.

A Acurácia é uma métrica que diz o quanto o modelo acertou em todas as previsões possíveis, ou seja, é a razão entre o somatório das previsões corretas (verdadeiros positivos mais verdadeiros negativos) sobre o somatório de todas as previsões. A equação abaixo mostra a forma de cálculo desta medida:

$$Acuracia = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.25}$$

O *Recall* mede a proporção de eventos que foram identificados corretamente, isto é, avalia o quão bom o modelo é para prever positivos considerando que a classe que se quer prever é a de positivos. Essa métrica é definida como a razão entre verdadeiros positivos sobre a soma de verdadeiros positivos com negativos falsos, conforme é observado na equação abaixo:

$$Recall = \frac{TP}{TP + FN} \tag{2.26}$$

3. Metodologia

3.1 Construção do banco de dados

Para a realização deste trabalho, os dados foram coletados do site oficial da NFL - National Football League, cujo endereço está presente nas referências bibliográficas [13]. Inicialmente, foram coletadas informações de todos os jogos da temporada regular (*Regular Seasson*) das temporadas de 2014 e 2015, e após o tratamento dos dados, retirando registros faltantes e observações incompletas, o banco de dados final utilizado para o desenvolvimento do trabalho ficou com 413 jogos registrados.

3.2 Variáveis analisadas

Foram consideradas para o estudo 38 variáveis relacionadas aos times envolvidos nas partidas. Tais variáveis representam ações de campo, scouts (informações quantitativas detalhadas sobre um time ou atleta) e características de jogo das equipes que podem ser quantificadas em números, traduzindo a qualidade do time.

Todas as estatísticas das equipes consideradas no banco de dados, são referentes às partidas realizadas na temporada até o jogo anterior ao que foi registrado, ou seja, as variáveis captam o nível da equipe no momento da partida observada.

Das 38 variáveis que foram utilizadas na construção do modelo, duas são dicotômicas e duas são discretas, sendo todas as outras contínuas. As duas variáveis dicotômicas representam o resultado das equipes mandante (que joga no seu próprio estádio) e visitante (que joga no estádio do adversário) na última partida realizada por elas, sendo que 1 representa a derrota e 0 a vitória. As duas variáveis discretas se referem ao número de vitórias de cada uma das duas equipes envolvidas na partida nos últimos três jogos realizados por elas, assim, essa variável varia apenas entre 0 e 3.

A variável resposta dicotômica Y é denotada como sendo 1 a vitória do time mandante e 0 a derrota. Para facilitar o desenvolvimento do modelo, as variáveis independentes referentes ao time mandante foram codificadas como x_i, e as variáveis referentes ao time visitante como z_i. Na Tabela 3.1, pode-se ver as variáveis preditoras e suas codificações.

Tabela 3.1: Variáveis preditoras e suas respectivas nomenclaturas.

Nomenclatura	Variável
x_1	% de vitórias na temporada
x_2	Número de vitórias nos últimos 3 jogos
x_3	Resultado do último jogo
x_4	Média de pontos marcados por partida
x_5	Média de jardas conquistadas por partida
x_6	Média de pontos cedidos por partida
x_7	Média de jardas cedidas por partida
x_8	Média de jardas de passe conquistadas por partida
x_9	Média de jardas de passe cedidas por partida
x_{10}	% de conversões em terceiras descidas do ataque
x_{11}	% de conversões em terceiras descidas da defesa
x_{12}	Média de sacks realizados por partida
x_{13}	Média de sacks sofridos por partida
x_{14}	Média de faltas feitas por jogo
x_{15}	Média de tempo de posse de bola por partida
x_{16}	Média de diferença de turnovers por partida
x_{17}	% de vitórias na temporada como mandante
x_{18}	Média de pontos marcados por partida como mandante
x_{19}	Média de pontos cedidos por partida como mandante
z_1	% de vitórias na temporada
z_2	Número de vitórias nos últimos 3 jogos
z_3	Resultado do último jogo
z_4	Média de pontos marcados por partida
z_5	Média de jardas conquistadas por partida
z_6	Média de pontos cedidos por partida
z_7	Média de jardas cedidas por partida
z_8	Média de jardas de passe conquistadas por partida
z_9	Média de jardas de passe cedidas por partida
z_{10}	% de conversões em terceiras descidas do ataque
z_{11}	% de conversões em terceiras descidas da defesa
z_{12}	Média de sacks realizados por partida
z_{13}	Média de sacks sofridos por partida
z_{14}	Média de faltas feitas por jogo
z_{15}	Média de tempo de posse de bola por partida
z_{16}	Média de diferença de turnovers por partida
z_{17}	% de vitórias na temporada como visitante
z_{18}	Média de pontos marcados por partida como visitante
z_{19}	Média de pontos cedidos por partida como visitante

3.3 Desenvolvimento do modelo

Para realizar a manipulação e tratamento do banco de dados, estimação do modelo de regressão, construção dos gráficos e demais análises complementares foi utilizado os softwares SAS Enterprise Guide 7.1, Microsoft Office Excel e R Studio.

Para a estimação do modelo de regressão logística binária foi incluído como variável resposta Y a coluna do banco de dados que representa o resultado da partida observada, sendo 1 a vitória

do time mandante e 0 a derrota. Foram incluídas inicialmente no modelo todas as 38 covariáveis citadas na subseção anterior.

Para o desenvolvimento do trabalho, a base de dados foi dividia em duas, sendo uma utilizada para a estimação do modelo (cerca de 75% dos dados) e a outra para teste (cerca de 25% dos dados). Foram utilizadas informações de 309 jogos registrados no banco de dados para a estimação do modelo, e após a obtenção do modelo final, utilizou-se 104 jogos para validar o poder preditivo do modelo.

Os parâmetros do modelo foram estimados através do método da máxima verossimilhança, utilizando a técnica iterativa de estimação de Newton-Raphson. No processo de seleção das variáveis, foi utilizado o método de eliminação Backward, e para atestar a significância do parâmetro no modelo foi utilizado o teste Wald. Após a obtenção das variáveis que compõem o modelo final e seus respectivos parâmetros, verificou-se se o critério de convergência do modelo foi satisfeito.

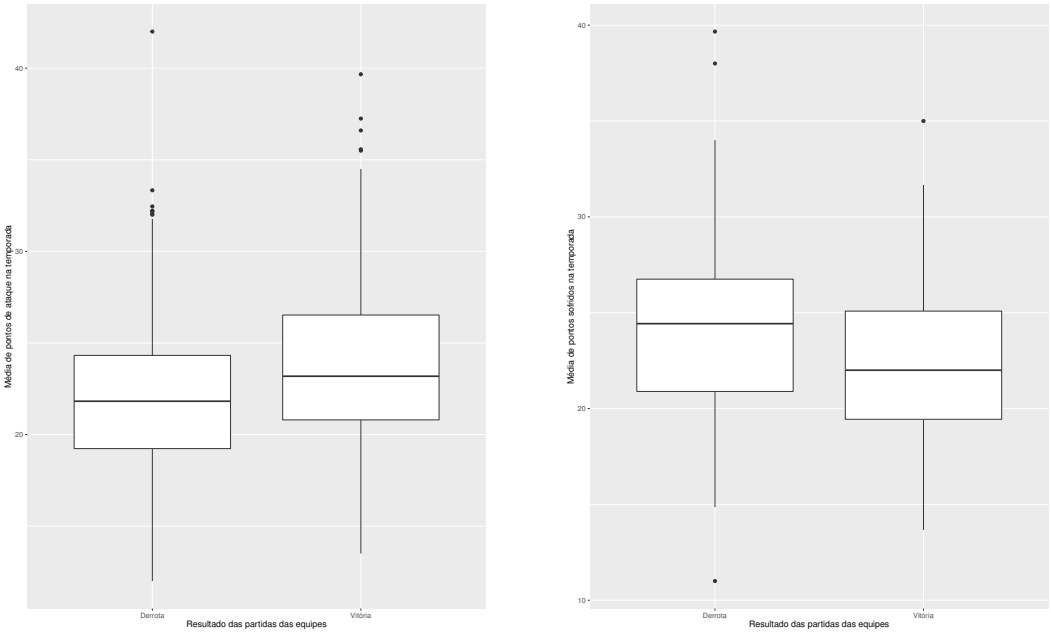
Além disso, estudou-se também a qualidade de ajuste do modelo através dos testes Deviance, Pearson e Hosmer-Lemeshow. Como parte importante do processo de análise do modelo ajustado, investigou-se visualmente alguns gráficos de diagnóstico, como a curva ROC, resíduos de Pearson e Deviance, Leverage e DF Beta. Por fim, foi construída a matriz de confusão do modelo ajustado, onde foi possível obter o poder preditivo do modelo em relação aos dados observados.

4. RESULTADOS

4.1 ESTATÍSTICAS DESCRITIVAS

Algumas variáveis consideradas importantes foram analisadas previamente para se obter informações e entender melhor o comportamento delas em relação ao contexto do estudo.

Por ser a medida final que define o vencedor de uma partida de futebol americano, a pontuação das equipes é uma variável importante que pode fornecer algumas informações complementares. Como era de se suspeitar, as equipes que vencem as partidas possuem geralmente uma média de pontos marcados maior que as equipes que perdem, por outro lado, os times que possuem médias de pontos sofridos maiores são os que mais perdem as partidas e os que possuem menores médias de pontos sofridos vencem mais. Tal resultado é ilustrado pela Figura 4.1.



(a) Média de pontos de ataque por resultado da equipe na partida.

(b) Média de pontos sofridos por resultado da equipe na partida.

Figura 4.1: Médias de pontos das equipes em cada partida.

Assim como na maiores dos esportes coletivos, o mando de campo é tido subjetivamente como fator determinante para o sucesso em um confronto. Na Figura 4.2, pode-se observar que na NFL não é diferente, as equipes que jogam em casa possuem uma frequência maior de

vitórias do que de derrotas, pois é no seu próprio estádio que o time recebe o apoio da torcida, onde os jogadores conhecem e estão acostumados com o ambiente e o clima local, entre outros fatores.

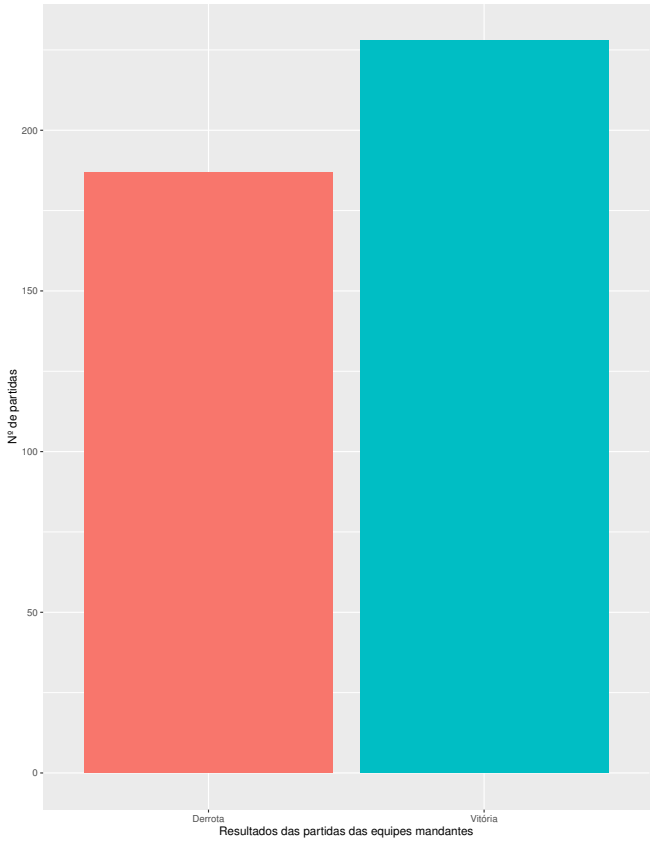


Figura 4.2: Média de pontos sofridos por resultado da equipe na partida.

O número de vitórias na temporada também é um fator chave para se definir a qualidade de uma equipe. Um time que está em boa fase na temporada, obviamente terá um aproveitamento alto, e consequentemente, estas equipes chegam na maioria das vezes como favoritas em confrontos contra equipes com aproveitamento mais baixo. Na Figura 4.3, tem-se que as equipes que vencem as partidas possuem aproveitamento de vitórias maior que as que perdem.

Outro scout importante que subjetivamente pode dizer muito sobre a qualidade e sucesso de uma equipe na NFL, é a diferença de Turnovers, que é a quantidade de bolas perdidas pelo time subtraída da quantidade de bolas recuperadas, ou seja, essa medida mostra o quão bem um time consegue roubar a bola do adversário e conseguir vantagem dentro da partida com isso. As equipes que vencem as partidas, possuem uma média de diferença de Turnovers na temporada positiva, ou seja, roubam mais a bola do adversário do que perdem, e o contrário acontece com as equipes que perdem as partidas, como é observado na Figura 4.4.

É senso comum dizer que as melhores equipes possuem ataque e defesa fortes, porém tal ideia não é totalmente verdade no caso da NFL. Na Figura 4.5, pode-se observar a relação entre a média de pontos de ataque e a média de pontos sofridos das equipes antes de cada partida realizada. Como pode-se perceber visualmente, a relação entre as duas variáveis é fraca, mostrando que poucas equipes possuem um bom desempenho conjunto de seu ataque e

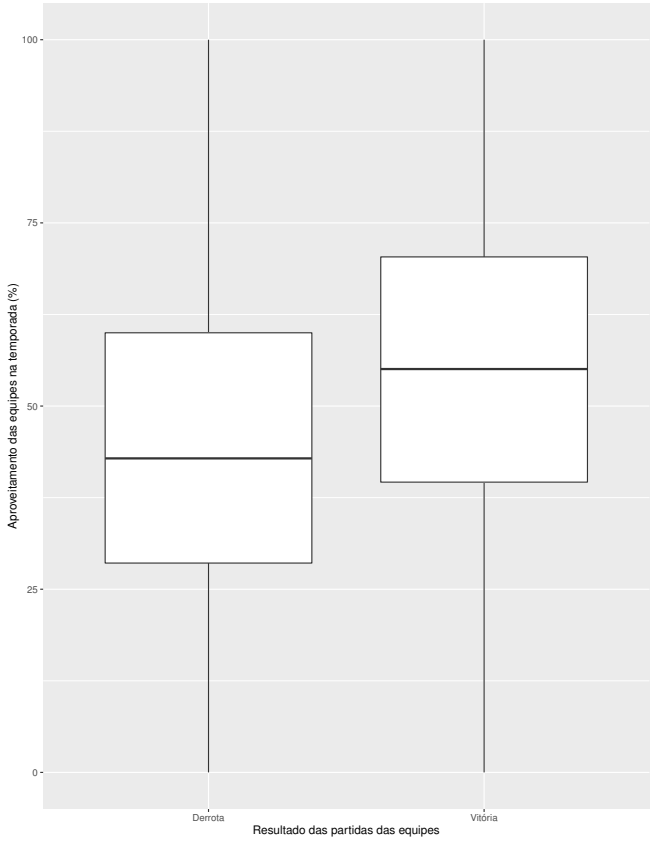


Figura 4.3: Aproveitamento na temporada (%) por resultado da equipe na partida.

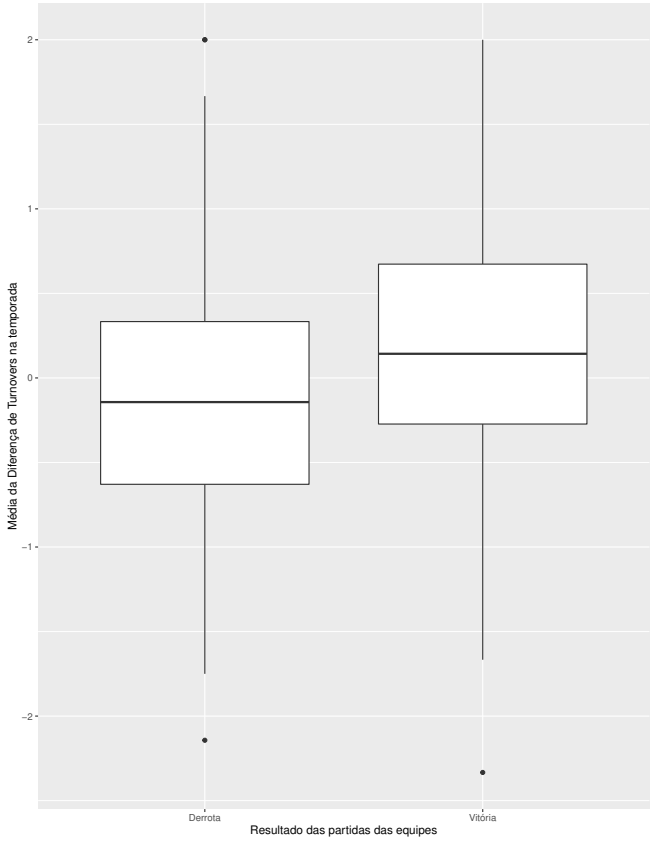


Figura 4.4: Média da diferença de Turnovers na temporada por resultado da equipe na partida.

defesa, e vice-versa. Tal resultado sugere que boas equipes podem ter um ataque forte e uma defesa não tão boa, ou podem ter um ataque não muito produtivo compensado por uma defesa sólida.

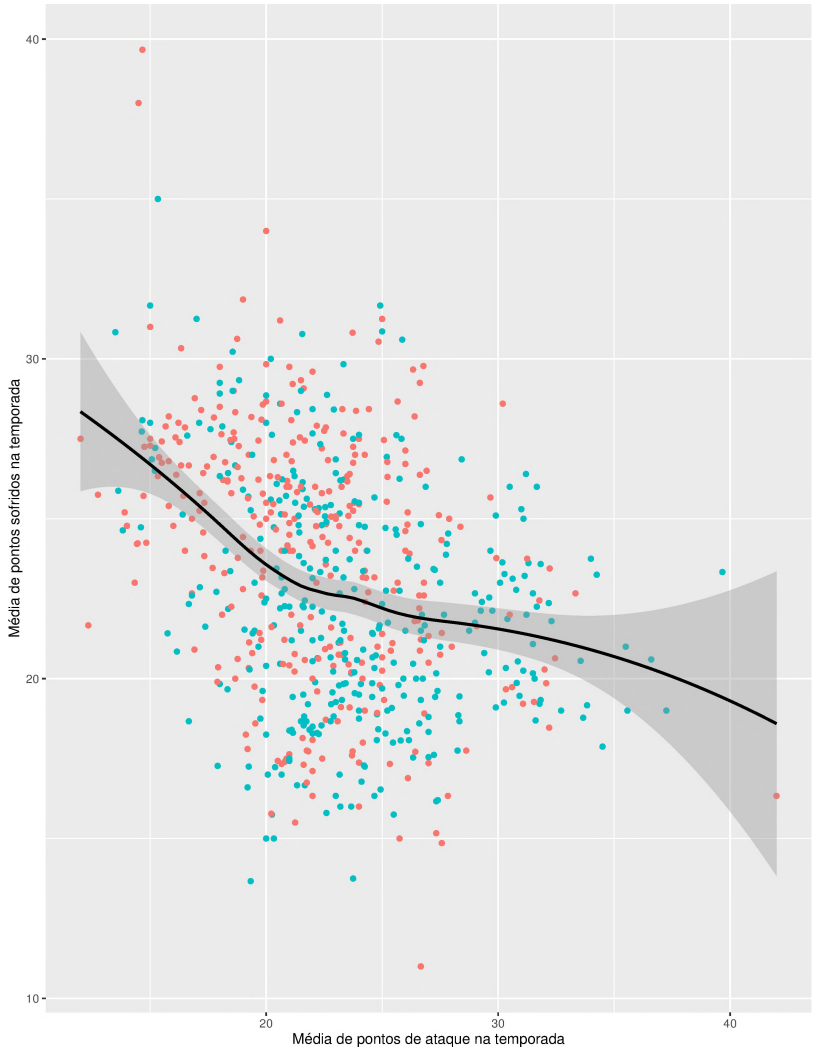


Figura 4.5: Dispersão entre a média de pontos marcados e sofridos pelas equipes nas partidas da temporada e a curva de tendência.

4.2 Modelo estimado

Para selecionar as variáveis que irão compor o modelo de regressão logístico foi utilizado o critério *Backward*, em que inicialmente ajustou-se o modelo com todas as variáveis independentes e retirou-se uma de cada vez. Como critério para retirar uma variável do modelo, foi utilizado o teste de Wald ao nível de 5% de significância, ou seja, as variáveis cujo parâmetro não foi significativo.

Ao fim do processo, permaneceram apenas 5 variáveis no modelo final, sendo elas x_4, x_8, x_{11}, z_1 e z_3, que correspondem respectivamente à média de pontos marcados por partida pela equipe mandante, média de jardas de passe conquistadas por partida da equipe mandante, conversões em terceiras descidas da defesa da equipe mandante (%), aproveitamento de vitórias

na temporada da equipe visitante (%) e o resultado do último jogo da equipe visitante. Os resultados do modelo estimado são apresentados na Tabela 4.1.

Tabela 4.1: Resultados do modelo final ajustado.

Parâmetros	G.L.	Estimado	Desvio Padrão	Qui-Quadrado Wald	p-valor
Intercepto	1	2,7604	1,4628	3,5608	0,0592
x_4	1	0,1559	0,0378	17,0459	<,0001
x_8	1	-0,0547	0,0253	4,6763	0,0306
x_{11}	1	-0,0126	0,0044	8,4389	0,0037
z_1	1	-0,0245	0,0062	15,6666	<,0001
z_3	1	0,5864	0,2832	4,2868	0,0384

Desta forma, após a estimação dos parâmetros, o preditor linear do modelo ajustado é dado pela expressão:

$$ln\left[\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right] = 2.7604 + 0.1559x_4 - 0.0547x_8 - 0.01226x_{11} - 0.0245z_1 + 0.5864x_3 \qquad (4.1)$$

4.3 INTERPRETAÇÃO DOS PARÂMETROS

Em regressão logística, podemos interpretar os parâmetros estimados do modelo através da Razão de chances (*Odds ratio*). Na Tabela 4.2, tem-se os valores estimados da razão de chances para cada parâmetro estimado e seus respectivos intervalos de confiança.

Tabela 4.2: Estimativas pontuais e intervalares das Odds Ratio.

Parâmetros	Odd Rate	Intervalo de confiança (95%)
x_4	1,169	1,085 - 1,258
x_8	0,947	0,901 - 0,995
x_{11}	0,987	0,979 - 0,996
z_1	0,976	0,964 - 0,988
z_3	1,798	1,032 - 3,132

De acordo com as Odds Ratio estimadas, pode-se dizer que a variável x_4 referente a média de pontos marcados pela equipe mandante tem impacto significativo nas chances de vitória do time da casa, pois quanto maior é a média de pontos marcados pela equipe maiores são as chances de vitória em uma partida, sendo que para cada unidade a mais na média de pontos, as chances de vitória aumentam em cerca de 17%. Na prática, uma equipe deve priorizar ter um ataque competitivo para aumentar as chances de sucesso na liga.

A Odd Ratio da variável x_8 mostra que o estilo de jogo interfere nas chances de vitória da equipe, pois quanto maior a média de jardas de passe conquistadas pela equipe mandante, menores as chances de vitória. Para cada unidade que se aumenta na média de jardas de passe, tem-se que as chances de vitória se reduz em cerca de 5,3% para a equipe da casa. Tal resultado evidencia que as equipes não devem focar apenas no jogo aéreo, e sim ter um bom desempenho

no jogo terrestre também, diminuindo consequentemente a média de jardas conquistadas pelo passe.

Observou-se também que a variável x_{11}, que se refere à quantidade de terceiras descidas cedidas pela defesa da equipe da casa, interfere nas chances desta equipe sair vitoriosa do confronto. De acordo com a Odds ratio estimada, tem-se que para cada unidade aumentada na quantidade de terceiras descidas cedidas (em %), as chances de vitória na partida diminuem em cerca de 1,3% para a equipe anfitriã. As jogadas de terceira descida da equipe visitante geralmente inflamam o estádio, onde a torcida local faz o máximo de barulho possível para atrapalhar e pressionar os jogadores adversários. Tal fato pode ser relacionado com o bom desempenho em casa de equipes que possuem torcidas "agressivas", dando sentido à significância da variável x_{11} no modelo estimado.

Em relação às varíáveis referentes ao time visitante, constatou-se que a variável z_1 referente ao desempenho da equipe durante a temporada, tem impacto significativo nas chances de vitória do time local. Para cada unidade acrescida no aproveitamento da equipe visitante (em %), as chances do time mandante vencer a partida diminuem cerca de 2,4%. De acordo com a Odds ratio estimada, tal resultado é um pouco obvio do ponto de vista lógico e esportivo, pois quanto maior o aproveitamento do time visitante maior sua qualidade, e consequentemente quanto maior sua qualidade mais difícil será para a equipe local vencer a partida.

Por fim, outra variável relacionada à equipe visitante que foi significativa no modelo é a z_3, que é uma variável dicotômica que informa o resultado da equipe visitante no jogo anterior, sendo 0 a vitória e 1 a derrota. Quando o time visitante é derrotado na última partida antes do confronto a ser previsto, as chances de vitória dos mandantes aumentam em 79,8%, mostrando o quanto uma derrota pode abalar e influenciar o desempenho de um time no jogo posterior.

4.4 Testes de ajuste do modelo

Nesta seção, serão apresentados os resultados de ajuste do modelo por meio dos testes de Resíduos de Pearson, Deviance e Hosmer-Lemeshow. Sempre que possível, é ideal analisar mais de um teste para avaliar a adequação do modelo, porém, para o caso deste trabalho, de acordo com a disposição do formato dos dados, o teste de Hosmer-Lemeshow é o mais adequado para a situação.

De acordo com a Tabela 4.3, percebe-se que o Teste de qualidade de ajuste de Pearson atestou o bom ajustamento dos dados ao modelo estimado. O p-valor do teste maior que o nível de significância estabelecido de 5% indica a não rejeição da hipótese nula de que o modelo ajustado é adequado.

O teste de Deviance foi significativo ao nível de confiança estabelecido indicando que o modelo não está bem ajustado. Para o formato da estrutura de dados utilizados na estimação do modelo, que está no modo binário resposta/frequência, o p-valor do teste de Deviance tende a ser menor, fazendo com que o resultado do teste da Deviance não seja de todo confiável para essa estrutura. Neste caso, os resultados do teste Hosmer-Lemeshow são mais confiáveis.

Além disso, para atestar a qualidade de ajuste devem ser utilizados sempre que possível, mais que um teste, sendo que neste caso, dois dos testes aceitaram a hipótese de bom ajuste do modelo.

Tabela 4.3: Teste de qualidade de ajuste de Pearson

Critério	Valor	G.L.	Valor/G.L.	p-valor
Deviance	386,82	302	1,2808	0,0007
Pearson	331,93	302	1,0991	0,1138

O teste de Hosmer-Lemeshow também indicou que o modelo obtido explica adequadamente os dados observados, ao nível de 5% de significância. Como é comumente feito na literatura, os dados foram divididos em 10 grupos de acordo com as probabilidades previstas. Percebe-se que o número de resposta observada e esperada são bem próximos em todos os grupos, indicando o bom ajuste do modelo, de acordo com a Tabela 4.4.

O resultado do teste é apresentado na Tabela 4.5, confirmando pelo p-valor o bom ajuste do modelo.

Tabela 4.4: Agrupamentos para o teste de Hosmer-Lemeshow

		Resultado = 1		Resultado = 0	
Grupo	Total	Observado	Esperado	Observado	Esperado
1	31	9	6,84	22	24,16
2	31	9	10,26	22	20,74
3	31	7	12,64	24	18,36
4	31	13	14,49	18	16,51
5	31	17	15,96	14	15,04
6	31	20	17,51	11	13,49
7	31	21	18,97	10	12,03
8	31	19	20,38	12	10,62
9	31	25	22,26	6	8,74
10	29	23	23,68	6	5,32

Tabela 4.5: Resultado do Teste Hosmer-Lemeshow

Qui-quadrado	G.L.	p-valor
8,7245	8	0,3661

4.5 Gráficos de diagnósticos e desempenho do modelo

A curva ROC (Receiver Operating Characteristic) está entre as métricas mais utilizadas para avaliação de um modelo de classificação, como o desenvolvido neste trabalho. A ROC mostra o quão bem um modelo criado pode diferenciar duas classificações, no caso deste trabalho, distinguir a vitória ou derrota de uma equipe.

A ROC possui dois parâmetros, que são eles a taxa de verdadeiros positivos, e a taxa de falsos positivos, traçando-os em diferentes limiares de classificação para cada ponto de corte, que no caso deste trabalho, foi definido como 0,33.

Para simplificar a análise da curva ROC, a AUC (Area Under the ROC Curve) é uma maneira de resumir a ROC em um único valor, que varia de 0 até 1, sendo que quanto maior o AUC, melhor a capacidade de classificação do modelo.

Para o modelo desenvolvido neste trabalho, como observado na Figura 4.6, obteve-se um AUC de 0,7185, tal valor que é considerado bom para um modelo de classificação. Pode-se dizer então, que o modelo logístico acerta corretamente 71,8% das predições feitas com a base de dados utilizada na estimação dos parâmetros do modelo.

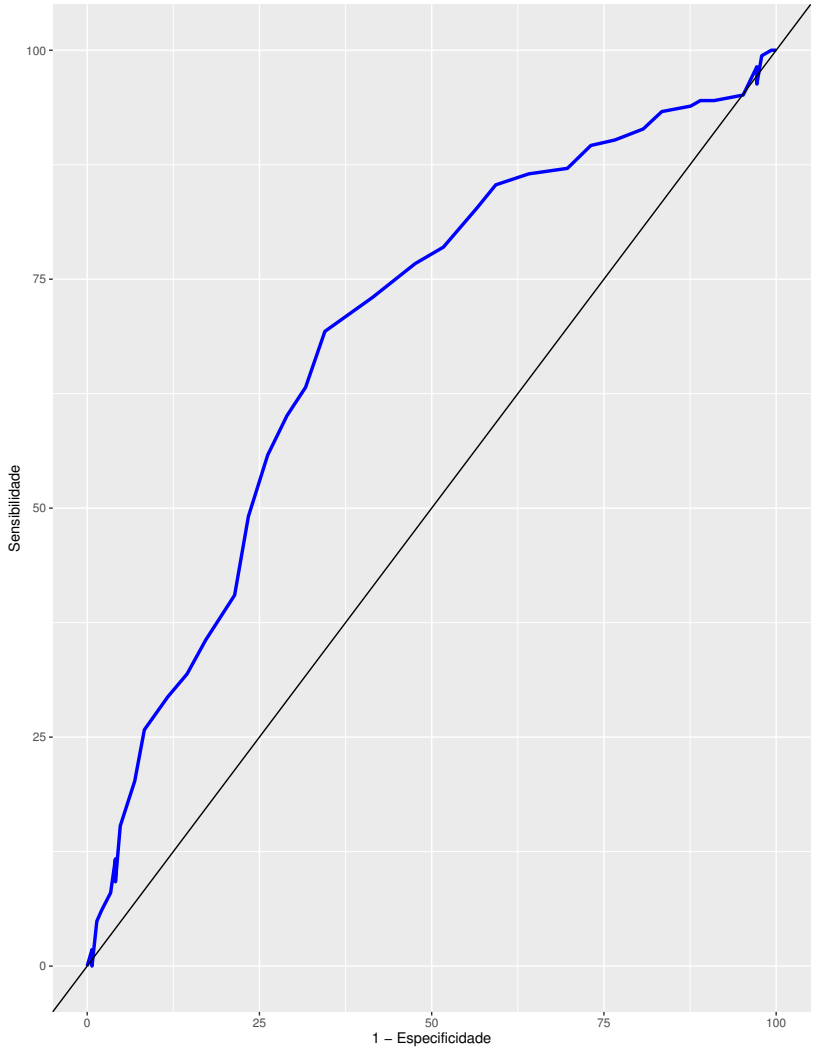


Figura 4.6: Curva ROC (Receiver Operating Characteristic)

Considerando o ponto de corte estimado de 0,33, para cada observação, foi calculado a probabilidade de vitória predita pelo modelo. Quando o valor predito foi maior que 0,33, considerou-se a predição de vitória para a equipe mandante e, caso contrário, a predição foi de derrota para aquela equipe.

Com os resultados de vitória e derrota observados e preditos pelo modelo, construiu-se a matriz de confusão. Foram reservadas cerca de 25% das observações (104 jogos) do banco de

dados para testar o poder preditivo do modelo.

O poder preditivo do modelo logístico estimado foi avaliado por meio de algumas métricas de desempenho que são baseadas na Matriz de confusão.

Tabela 4.6: Matriz de confusão do modelo logístico estimado.

	Resultados observados	
Resultados preditos	Vitória	Derrota
Vitória	55	32
Derrota	9	8

A partir da Matriz de confusão, disposta na Tabela 4.6, foram calculadas a Acurácia e *Recall* das predições realizadas pelo modelo com a base de dados teste. Os resultados dessas métricas, conforme se observa na Tabela 4.7, mostram que a capacidade preditiva do modelo é aceitável.

A Acurácia das predições realizadas é de 60,58%, um valor considerado aceitável para o modelo preditivo estimado neste trabalho, visto que é difícil prever eventos deste tipo e que a amostra utilizada para teste é pequena. Resultado semelhante pode ser visto em dados divulgados pela ESPN (*Entertainment and Sports Programming Network*), emissora especializada em eventos esportivos, principalmente em esportes americanos, que possui em seu site uma seção onde é realizada predições de jogos da NFL de forma probabilística [6]. Ao avaliar a assertividade do modelo da ESPN em jogos passados, observou-se que cerca de 71% dos jogos são preditos corretamente pelo modelo preditivo deles (considerando ponto de corte de 0,5). Levando em conta que eles possuem um histórico de jogos bem maior que o utilizado neste trabalho, além de aparatos estatísticos mais robustos para estimação de modelos, a Acurácia obtida para este modelo não é ruim, pois é cerca de 10% menor que o modelo referência da ESPN.

Ao avaliar o quão bom o modelo estimado é na predição de vitórias da equipe mandante, observa-se um desempenho melhor que na predição geral, visto que o *Recall* das predições realizadas pelo modelo é de 85,94%, sendo este um resultado muito bom para um modelo preditivo considerando o contexto dos dados, isto é, de todos os jogos que o modelo prevê vitória da equipe mandante, cerca de 86% dessas predições estarão corretas.

Tabela 4.7: Resultado das métricas de avaliação das predições realizadas pelo modelo.

Métrica	Resultado
Acurácia	60,58%
Recall	85,94%

Para identificar tais observações, é utilizado o gráfico com a medida Leverage de cada observação. Pontos mais afastado devem ser avaliados com cautela. Para este trabalho, de acordo com o tamanho da amostra e número de parâmetros do modelo, valores de Leverage maiores que 0,0389 devem ser investigados.

Conforme é apresentado na Figura 4.7, percebe-se alguns possíveis pontos de alavanca, que foram avaliados posteriormente.

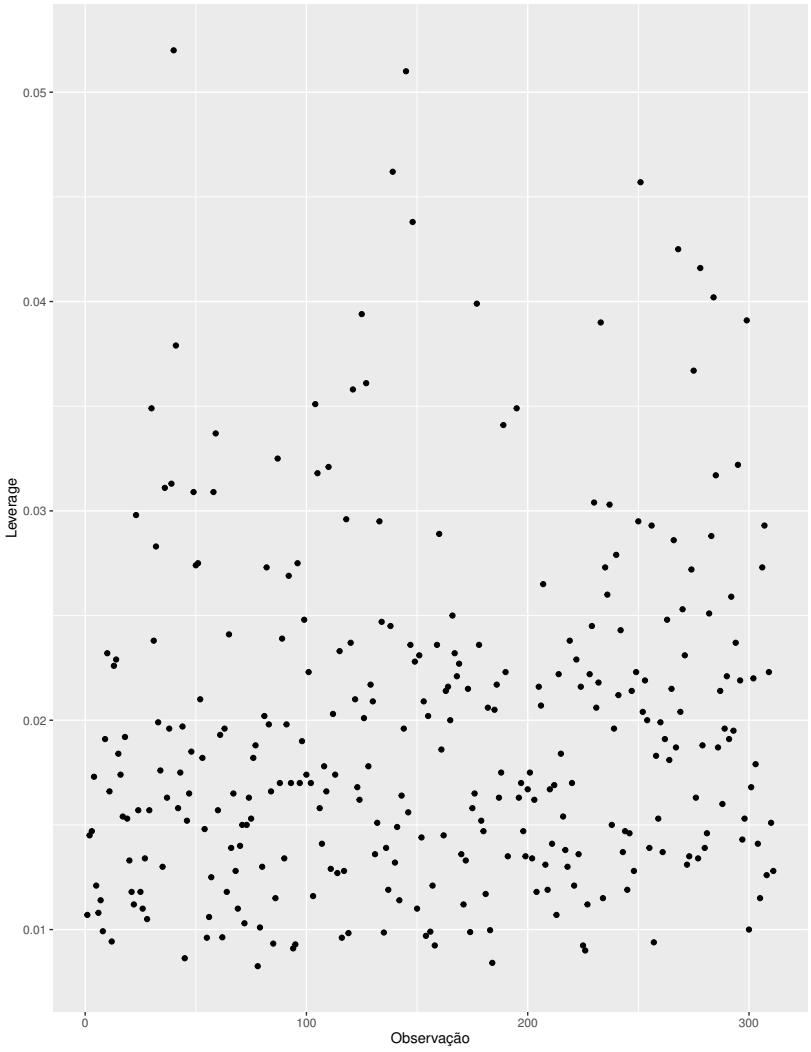


Figura 4.7: Gráfico de diagnóstico Leverage contra Observações

Um dos diagnóstico mais utilizados na avaliação da qualidade de ajuste de um modelo, são os gráficos de resíduos de Pearson e Deviance, sendo possível verificar visualmente a heterogeneidade das variâncias. O padrão para esse tipo de gráfico é uma distribuição aleatória de média 0 e amplitude constante.

Na Figura 4.8, pode-se ver que os resíduos de Pearson e Deviance estão aleatoriamente dispersos em torno do eixo 0, e não apresentam nenhuma assimetria em sua distribuição, indicando portanto a homogeneidade dos resíduos. Porém, é possível identificar no gráfico de resíduos de Pearson um ponto muito discrepante entre as observações 0 e 50, que foi investigado posteriormente como um possível ponto de influência na estimação dos parâmetros do modelo.

Esse ponto em especial, é referente a um jogo do início da temporada, onde uma das equipes possuía uma média muito alta de pontos marcados. Decidiu-se em manter essa observação, por não se tratar de um caso atípico, visto que podem haver durante a história, equipes muito superiores com resultados muito acima das demais, e tal informação é importante para a estimação dos parâmetros do modelo, visto que o que se deseja é realmente captar informações que retratam a qualidade da equipe. Observou-se também outros pontos acima do valor referência

dos gráficos, que foram investigados posteriormente.

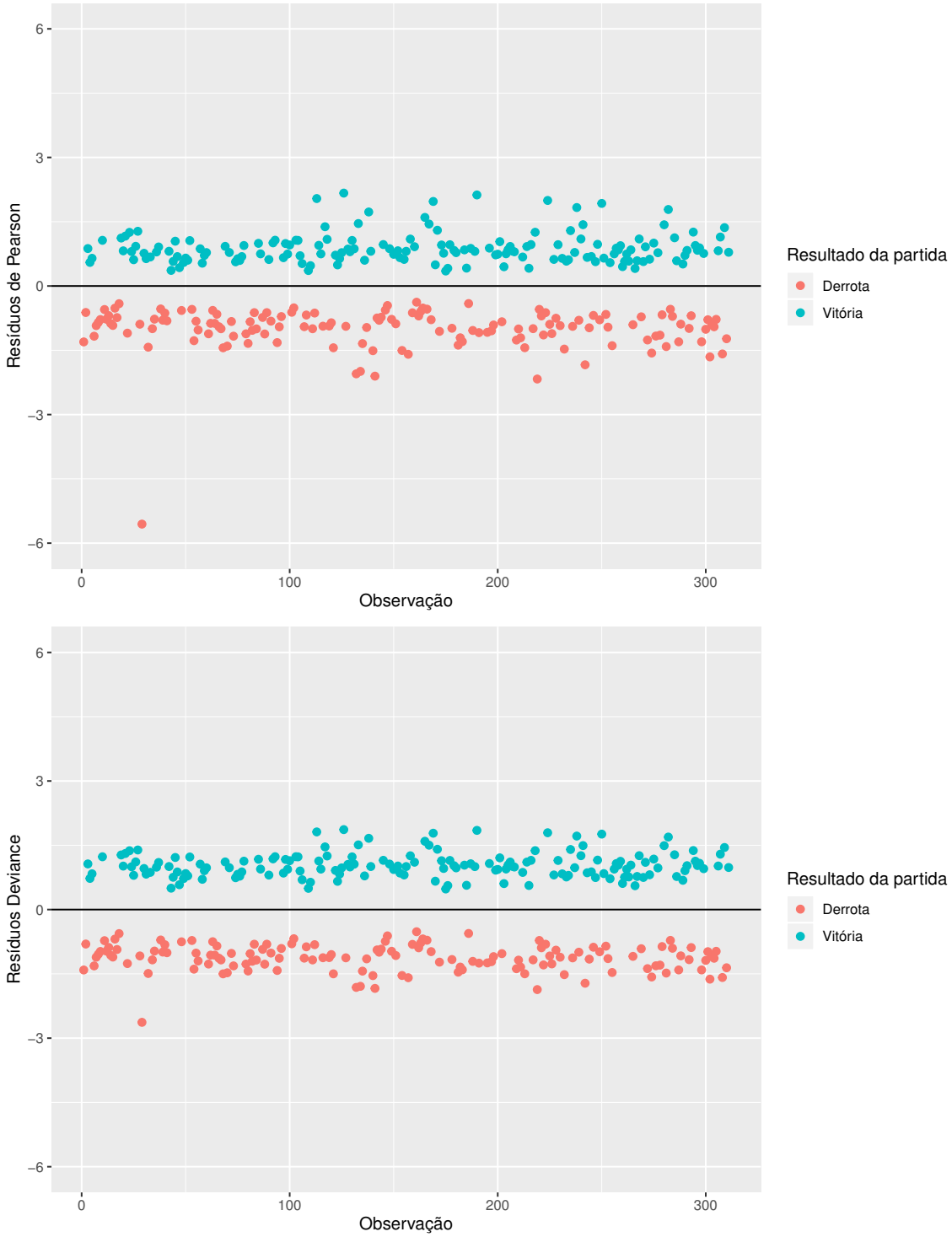


Figura 4.8: Gráfico de diagnóstico dos resíduos de Pearson e Deviance contra Observações.

Um ponto é influente se sua exclusão do ajuste de regressão causa uma mudança significativa no modelo estimado. Uma técnica desenvolvida para identificar essas observações é a medida DFBETA, que mede a influência da observação i sobre o coeficiente de X_j.

Um valor alto para a medida DFBETA indica que a observação i tem influencia significativa na estimativa do coeficiente angular da variável explicativa X_j. Para o presente trabalho, valores de DFBETA maiores que 0,1138 devem ser investigados, pois podem ser possíveis pontos de

influência na estimação dos parâmetros.

Como é observado na Figura 4.9, os valores de DFBETA para todos os parâmetros do modelo estão dispersos e distribuídos aleatoriamente em torno do 0, porém há alguns pontos acima do valor referência. Alguns desses pontos foram investigados posteriormente.

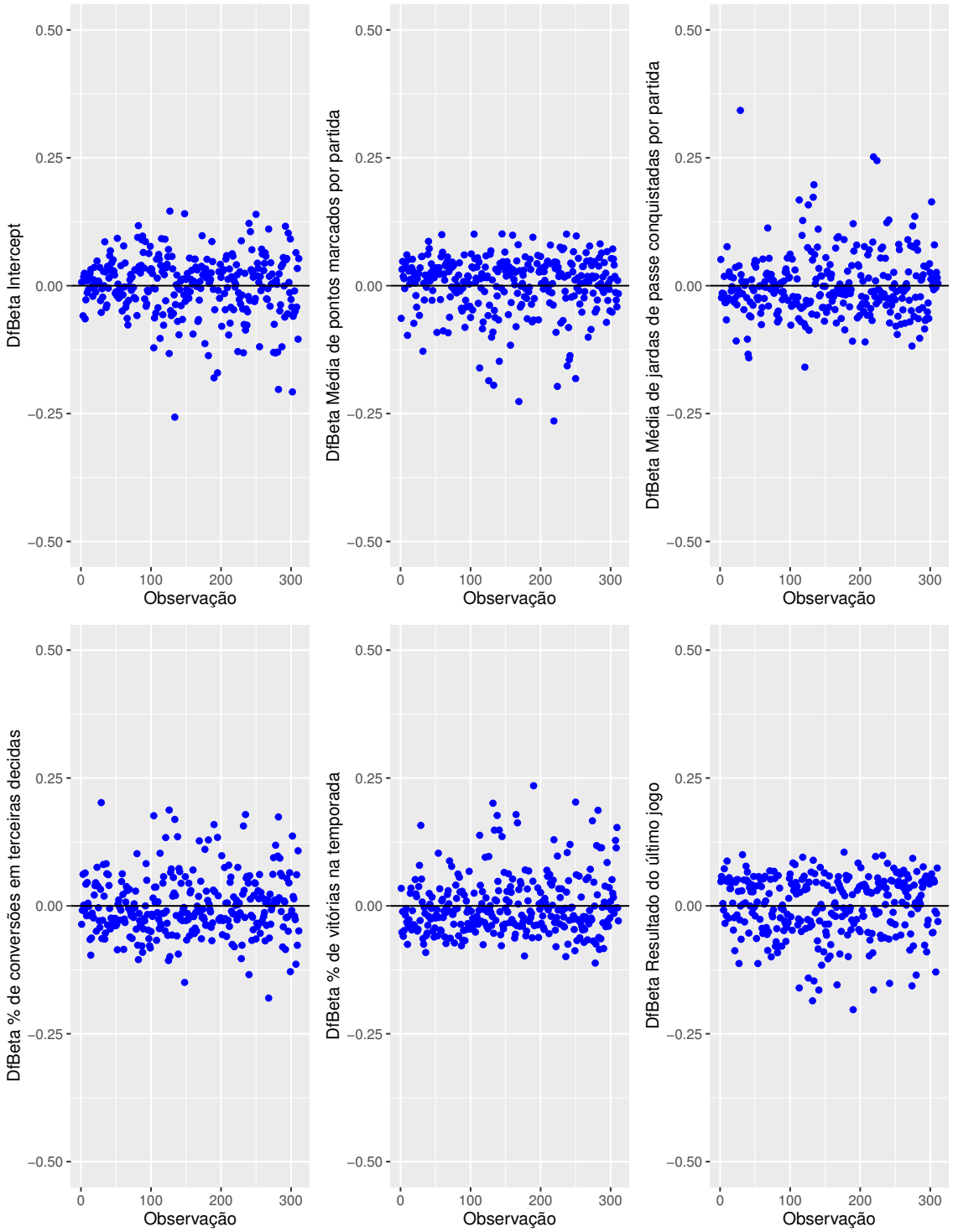


Figura 4.9: Gráfico de diagnóstico DFBETA vs Observações de todos os parâmetros do modelo.

De acordo com a maioria dos gráficos de diagnóstico, há algumas observações que precisam ser investigadas como possíveis pontos que atrapalham a estimação dos parâmetros do modelo. Investigando a fundo tais registros, observou-se que se tratam sempre de times que possuíam re-

sultados muito bons, bem acima da média das demais equipes, isto é, são observações plausíveis para o cenário esportivo estudado e importantes na estimação do modelo.

Visto que esses registros "duvidosos"não foram raros e pontuais (há um número considerável deles), e que eles refletem a realidade da população por carregarem informações importantes sobre a qualidade de uma equipe e seu resultado em uma partida, que é justamente o que se deseja estudar, optou-se por não retirar esses registros do conjunto de dados para estimar outro modelo, a fim de eliminar o risco de se estimar um modelo que não reflete a realidade.

Outro ponto importante que deve ser considerado, é que na estimação de um modelo preditivo, as análise e ponderações feitas devem se basear em vários testes de diagnóstico e não em apenas alguns, pois dificilmente um modelo estimado para dados complexos e reais terá todos os diagnósticos perfeitos. Visto que o Futebol Americano é um dos esportes mais complexos do mundo, onde vários fatores podem interferir nas chances de vitória ou derrota de uma equipe, e que alguns testes atestaram a qualidade do ajuste do modelo aos dados, decidiu-se manter o modelo estimado inicialmente.

O resultado obtido aqui é comparável ao estudo de Barbieri (2012) [3], que propõe um método de estimação robusta para modelos de regressão logística utilizando os softwares R e SAS. Após observar e analisar se as observações discrepantes são plausíveis no contexto estudado, é utilizada algumas funções para estimar um novo modelo atribuindo pesos menores para as observações destoantes. Este recurso pode ser utilizado em trabalhos futuros na tentativa de verificar se tal método pode melhorar o modelo obtido no presente estudo.

Outra alternativa que pode ser utilizada também em trabalhos futuros para melhorar os resultados de diagnóstico do modelo, é aumentar o tamanho da amostra e acrescentar novas variáveis mensuráveis relacionadas as equipes e condições de jogo, agregando mais temporadas no banco de dados dos jogos. Possivelmente, isso pode ajudar no processo de estimação de um novo modelo.

Os testes de diagnóstico em geral mostraram um bom ajuste do modelo aos dados considerados. A capacidade preditiva do modelo teve ótimo desempenho na predição de vitórias, porém, na predição geral dos resultados não se mostrou tão eficiente quanto se espera de um modelo preditivo. Em trabalhos futuros este modelo pode ser melhorado aumentando a amostra de jogos utilizados no banco de dados e acrescentando algumas variáveis que não foram consideradas neste trabalho para estimação do modelo. Ainda assim, é importante salientar que o Futebol Americano é um esporte de extrema complexidade, onde diversos fatores que não podem ser quantificados podem influenciar no resultado de uma partida.

5. Conclusões

Neste trabalho foi possível estudar com êxito a relação entre algumas ações e estatísticas de times da NFL e a probabilidade de saírem vencedores de uma partida, através da regressão logística. No modelo ajustado, apenas 5 covariáveis foram significativas.

Verificou-se que a variável x_4 se relaciona positivamente com a probabilidade de vitória da equipe mandante, isto é, quanto maior a média de pontos marcados por partida, maior a chance de vitória. As covariáveis x_8, x_{11} e z_1 se relacionam negativamente com a probabilidade de vitória da equipe da casa, isto é, quanto maiores a média de jardas de passe conquistadas por partida pela equipe mandante, a porcentagem de conversões em terceiras descidas da defesa da equipe da casa, e o aproveitamento de vitórias na temporada da equipe visitante, menor a chance de vitória do time mandante. A variável dicotômica z_3, resultado do último jogo do visitante, também se relaciona de forma positiva com a probabilidade de vitória dos mandantes, isto é, se a equipe visitante perdeu o último jogo $z_3 = 1$, maior a chance de vitória da equipe mandante.

Referências Bibliográficas

[1] Action, P.: *Análise de regressão - Predição*, 2019. http://www.portalaction.com.br/analise-de-regressao/45-predicao, acessado em 07/06/2019.

[2] Adriano K. Suzuki, Luis E. B. Salasar, J. G. L. e. F. N. L.: *A Bayesian approach for predicting match outcomes: the 2006 (Association) Football World Cup*. Journal of the Operational Research Society, 61(1):1530–1539, 2010. https://doi.org/10.1057/jors.2009.127.

[3] Barbieri, N. B.: *Estimação Robusta para o Modelo de Regressão Logística*. Trabalho de Conclusão de Curso, 2012.

[4] Douglas Montgomery, G. G. V. e Peck, E. A.: *Introduction to Linear Regression Analysis*. John Wiley Sons, 5ª ed., 2012.

[5] Dyte, D. e Clarke, S. R.: *A ratings based Poisson model for World Cup soccer simulation*. Journal of the Operational Research Society, 51(1):993–998, 2000. https://doi.org/10.1057/palgrave.jors.2600997.

[6] ESPN: *NFL - Notícias, Times, Resultados, Estatísticas e Classificação*, 2019. https://www.espn.com.br/nfl/, acessado em 25/06/2019.

[7] Fontes, J.: *Invista em futebol*. Gente, 1ª ed., 2015.

[8] Hosmer, D. W. e Lemeshow, S.: *Applied Logistic Regression*. Wiley, 1ª ed., 2000.

[9] John Neter, Michael H. Kutner, C. J. N. e Wasserman, W.: *Applied Linear Statistical Models*. Irwin, 3ª ed., 1996.

[10] Karlis, D. e Ntzoufras, L.: *Analysis of sports data by using bivariate Poisson models*. The Statistician, 52(3):381–398, 2003. http://www2.stat-athens.aueb.gr/~jbn/papers2/08_Karlis_Ntzoufras_2003_RSSD.pdf.

[11] Lee, A. J.: *Modeling Scores in the Premier League: Is Manchester United really the Best?* Chance, 10(1):15–19, 1997. https://doi.org/10.1080/09332480.1997.10554791.

[12] Minitab: *Interpretar os principais resultados para Ajustar modelo logístico binário.*, 2019. https://support.minitab.com/pt-br/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-binary-logistic-model/interpret-the-

results/key-results/#step-3-determine-how-well-the-model-fits-your-data, acessado em 08/06/2019.

[13] NFL: *Official Site of the National Football League*, 2019. https://www.nfl.com/, acessado em 28/08/2018.

[14] Norman, J.: *Football still American's favorite sport to watch*, 2018. https://news.gallup.com/poll/224864/football-americans-favorite-sport-watch.aspx, acessado em 08/09/2018.

[15] Pregibon, D.: *Logistic Regression Diagnostic*. Ann Statist, 9(4):705–724, 1981. https://www.jstor.org/stable/2240841.