UNIVERSIDADE FEDERAL DE SÃO CARLOS
PROGRAMA INTERINSTITUCIONAL DE PÓS GRADUAÇÃO
EM ESTATÍSTICA

Thiago Souza de Melo

MODELOS DE REGRESSÃO PARA PLACARES DE JOGOS DIVIDIDOS EM *SETS*

São Carlos - SP

2026

Thiago Souza de Melo

MODELOS DE REGRESSÃO PARA PLACARES DE JOGOS DIVIDIDOS EM *SETS*

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Gustavo Henrique de Araújo Pereira
Coorientador: Prof. Dr. Jeremias da Silva Leão

USP – São Carlos
Janeiro de 2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Souza de Melo, Thiago
S729m MODELOS DE REGRESSÃO PARA PLACARES DE JOGOS DIVIDIDOS EM SETS / Thiago Souza de Melo; orientador Gustavo Henrique de Araújo Pereira; coorientador Jeremias da Silva Leão. -- São Carlos, 2025.
75 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Cadeias de Markov. 2. distribuição de Bernoulli. 3. jogos esportivos divididos em sets. 4. modelos de regressão bivariados. I. Henrique de Araújo Pereira, Gustavo , orient. II. da Silva Leão, Jeremias , coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Thiago Souza de Melo

REGRESSION MODELS FOR SCORES OF MATCHES DIVIDED INTO SETS

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Gustavo Henrique de Araújo Pereira
Co-advisor: Prof. Dr. Jeremias da Silva Leão

USP – São Carlos
January 2026

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Thiago Souza de Melo, realizada em 16/12/2025.

Comissão Julgadora:

Prof. Dr. Gustavo Henrique de Araujo Pereira (UFSCar)

Profa. Dra. Denise Aparecida Botter (USP)

Profa. Dra. Clarice Garcia Borges Demétrio (ESALQ/USP)

Prof. Dr. Rinaldo Artes (Insper)

Prof. Dr. Manoel Ferreira dos Santos Neto (UFC)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Primeiramente, agradeço a Deus pelo dom da vida e pelas forças concedidas ao longo desta caminhada.

À minha esposa e companheira de vida, pelo suporte, auxílio e fortalecimento constantes durante todos esses anos, fundamentais para a conclusão desta etapa.

Aos meus amigos e colegas de turma, que contribuíram de diferentes formas, seja nos momentos de estudo em grupo ou individuais, compartilhando conhecimentos, experiências e apoio mútuo.

Ao meu orientador, Professor Gustavo Henrique, por aceitar a orientação deste trabalho, pelos valiosos ensinamentos transmitidos, pela presença constante, disponibilidade para ouvir e atenção dedicada sempre que necessário. As reuniões semanais foram fundamentais para o meu desenvolvimento acadêmico e científico. Agradeço, ainda, por ser uma excelente pessoa e um profissional exemplar, tornando a trajetória acadêmica mais leve e enriquecedora.

Aos membros da banca examinadora, Profa. Dra. Denise Aparecida Botter, Profa. Dra. Clarice Garcia Borges Demétrio, Prof. Dr. Rinaldo Artes e Prof. Dr. Manoel Ferreira dos Santos Neto, pela disponibilidade, bem como pelas correções e sugestões que contribuíram significativamente para o aprimoramento deste trabalho.

À Universidade Federal do Amazonas (UFAM), pela concessão da licença para capacitação, que possibilitou a dedicação necessária para a conclusão desta formação ao longo desses quatro anos.

RESUMO

MELO, T. S. **MODELOS DE REGRESSÃO PARA PLACARES DE JOGOS DIVIDIDOS EM *SETS***. 2026. 79 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

O principal objetivo deste trabalho é propor modelos de regressão para o placar de jogos divididos em *sets*. Inicialmente, assumimos que os *sets* são independentes e obtemos a distribuição de probabilidade bivariada associada a essa proposta. Em seguida, propomos um modelo de regressão assumindo essa distribuição de probabilidade para a variável resposta. São discutidos aspectos inferenciais e de diagnóstico desse modelo, avalia-se a performance do estimador de máxima verossimilhança e são realizadas duas aplicações a dados reais. Posteriormente, assumimos dependência Markoviana entre os *sets* e, fixado o número de *sets* necessários para a vitória, obtemos a distribuição de probabilidade bivariada associada a essa proposta. Para finalizar o trabalho, propomos um modelo de regressão assumindo essa distribuição de probabilidade para a variável resposta. Assim como no modelo anterior, discutimos aspectos inferenciais e de diagnóstico, avaliamos a performance do estimador de máxima verossimilhança no modelo e realizamos uma aplicação a dados reais. Essa última aplicação, utilizando dados de Grand Slam de tênis, sugere que o modelo proposto é adequado para modelar os placares dessas partidas em função de covariáveis.

Palavras-chave: Cadeias de Markov, distribuição de Bernoulli, jogos esportivos divididos em sets, modelos de regressão bivariados.

ABSTRACT

MELO, T. S. **REGRESSION MODELS FOR SCORES OF MATCHES DIVIDED INTO SETS**. 2026. 79 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

The main objective of this work is to propose regression models for the score of games divided into sets. Initially, we assume that the sets are independent and obtain the bivariate probability distribution associated with this proposal. Then, we propose a regression model assuming this probability distribution for the response variable. Inferential and diagnostic aspects of this model are discussed, the performance of the maximum likelihood estimator is evaluated, and two applications to real data are carried out. Subsequently, we assume Markovian dependence between the sets and, fixing the number of sets required for victory, we obtain the bivariate probability distribution associated with this proposal. To conclude the work, we propose a regression model assuming this probability distribution for the response variable. As in the previous model, we discuss inferential and diagnostic aspects, evaluate the performance of the maximum likelihood estimator in the model, and perform an application to real data. This final application, using Grand Slam tennis data, suggests that the proposed model is appropriate for modeling the scores of these matches as a function of covariates.

Keywords: Markov chains, Bernoulli distribution, sports games divided into sets, bivariate regression models.

LISTA DE ILUSTRAÇÕES

Figura 2.1 – Gráfico normal de probabilidade com envelope simulado para o modelo ajustado com os dados da Copa Davis 39
Figura 2.2 – Gráfico normal de probabilidade com envelope simulado do modelo ajustado com os dados de tênis de mesa 42
Figura 4.1 – Gráfico de probabilidade normal com envelope simulado dos modelos . . . 61
Figura 4.2 – Gráfico de afastamento de verossimilhança do modelo 62

LISTA DE ALGORITMOS

Algoritmo 1 – Algoritmo para gerar valores pseudo-aleatórios. 48

LISTA DE TABELAS

Tabela 2.1 – Exemplo: Jogador A vence a partida em quatro *sets*. 24
Tabela 2.2 – Probabilidades para um jogo na versão melhor de cinco *sets*. 24
Tabela 2.3 – Derivadas de primeira e segunda ordem da função de ligação em relação aos parâmetros $\boldsymbol{\beta}$. 29
Tabela 2.4 – Estimativas do viés e REQM (em parênteses) dos estimadores dos parâmetros nos dados simulados do modelo considerando *sets* independentes. 34
Tabela 2.5 – Observações das eliminatórias da Copa Davis 2024. 36
Tabela 2.6 – Distribuição de frequências para os resultados dos placares da Copa Davis. . 37
Tabela 2.7 – Coeficientes de correlação de Pearson das variáveis transformadas dos dados de tênis da Copa Davis. 38
Tabela 2.8 – Critérios de seleção de modelos ajustados com dados de tênis da Copa Davis. 38
Tabela 2.9 – Estimativas de máxima verossimilhança do modelo ajustado com dados de tênis da Copa Davis. 38
Tabela 2.10–Observações iniciais e finais do Campeonato Mundial de Tênis de Mesa 2023. 40
Tabela 2.11–Distribuição de frequências para o resultado dos placares dos dados de tênis de mesa. 41
Tabela 2.12–Coeficientes de correlação de Pearson das variáveis transformadas dos dados de tênis de mesa. 41
Tabela 2.13–Estimativas de máxima verossimilhança do modelo ajustado com os dados de tênis de mesa. 42
Tabela 3.1 – Evento: jogador B vencer a partida em três *sets*. 46
Tabela 3.2 – Evento: jogador A vencer a partida em cinco *sets*. 46
Tabela 3.3 – Evento: jogador A vencer a partida em seis *sets*. 47
Tabela 3.4 – Estimativa dos viés e REQM (em parênteses) dos estimadores dos parâmetros do modelo proposto para os dados simulados com alguns valores de n e $\boldsymbol{\theta} = (p_0, p_{00}, p_{11})$. 51
Tabela 4.1 – Estimativas do viés e REQM (em parênteses) dos estimadores do modelo considerando dependência entre os *sets* com os dados simulados. 58
Tabela 4.2 – Primeiras observações dos dados: Torneios de Grand Slam 2025. 59
Tabela 4.3 – Critérios de seleção de modelos. 60
Tabela 4.4 – Estimativas do modelo de regressão para os dado torneios de Grand Slam. . 61

Tabela 4.5 – RCs (em %) nas estimativas e erros padrões do modelo correspondente para os casos removidos indicados e valores de p dos dados do torneios de Grand Slam. 63
Tabela 4.6 – Distribuição de frequências dos resultados dos placares e previsões dos modelos. 64

SUMÁRIO

1 INTRODUÇÃO 19
1.1 Objetivos e organização da tese 20

2 MODELOS DE REGRESSÃO CONSIDERANDO *SETS* INDEPENDENTES 23
2.1 Distribuição de probabilidade bivariada proposta 24
2.1.1 Estimador de máxima verossimilhança *26*
2.1.2 Informação de Fisher *26*
2.2 Modelos de regressão 27
2.2.1 Estimação dos parâmetros *28*
2.3 Intervalos de confiança e testes de hipóteses assintóticos 30
2.4 Análise de diagnóstico 31
2.5 Estudos de simulação 32
2.6 Aplicações 34
2.6.1 Dados de tênis: Copa Davis *34*
2.6.2 Dados de tênis de mesa *39*

3 DISTRIBUIÇÃO DE PROBABILIDADE CONSIDERANDO DEPENDÊNCIA ENTRE OS *SETS* 45
3.1 Distribuição de probabilidade bivariada proposta 46
3.1.1 Estimação dos parâmetros *48*
3.1.2 Estudos de simulação *49*

4 MODELOS DE REGRESSÃO CONSIDERANDO DEPENDÊNCIA ENTRE OS *SETS* 53
4.1 Estimação dos parâmetros 54
4.2 Análise de diagnóstico 55
4.2.1 Análise de resíduos *55*
4.2.2 Influência global *56*
4.3 Estudos de simulação 56
4.4 Aplicação em dados reais 58

5 CONCLUSÕES 65
5.0.1 Trabalhos futuros *66*

REFERÊNCIAS 67

APÊNDICE A PROVAS E RESULTADOS 71
A.1 Cálculo do valor esperado da distribuição marginal 71
A.2 Cálculo do vetor escore e da matriz de informação de Fisher 72
A.2.1 Obtenção da função escore para o modelo *72*
A.2.2 Obtenção da matriz de informação de Fisher *74*

CAPÍTULO

1

INTRODUÇÃO

A modelagem estatística aplicada aos esportes é uma ferramenta poderosa que desempenha um papel fundamental na previsão de resultados, análise do desempenho de jogadores e equipes, e, principalmente, na tomada de decisões estratégicas. Nos últimos anos, a construção de modelos estatísticos para prever resultados de competições esportivas tem recebido muita atenção. Um livro que se destacou na popularização da análise de dados no contexto esportivo é Lewis (2004), que usou análise estatística para construir um time competitivo mesmo com um orçamento limitado. Além disso, em Thorn e Palmer (2009), é apresentada uma visão geral da história e do desenvolvimento da análise estatística em dados de esportes.

O esporte que tem sido objeto de inúmeros estudos envolvendo a modelagem estatística é o futebol, amplamente considerado o esporte profissional mais famoso no mundo. Existem diversos trabalhos na literatura que propõem modelos com o objetivo de predição de resultados em partidas de futebol. Por exemplo, Suzuki (2007) empregou a distribuição de Poisson bivariada, e levou em consideração uma variedade de variáveis, como estatísticas de equipes e jogadores, histórico de confrontos, condições de jogo e outros fatores relevantes para melhorar a precisão das previsões.

As modalidades esportivas disputadas em *sets*, como tênis e vôlei, têm atraído crescente interesse em estudos que envolvem modelagem estatística. A literatura acadêmica apresenta diversos trabalhos nessa área, com foco inicial em modelagens puramente estocásticas para calcular as probabilidades de vitória em pontos, *games*, *sets* e partidas. Essas modelagens não consideram variáveis explicativas na previsão dos resultados. A maioria dos modelos matemáticos quantitativos descreve essas probabilidades usando abordagens hierárquicas e cadeias de Markov (NEWTON; KELLER, 2005; O'MALLEY, 2008; CARRARI *et al.*, 2017; SIM; CHOI, 2020).

Alguns estudos exploram a combinação de modelos de regressão com cadeias de Markov para aproveitar as vantagens de ambas as abordagens. Esses modelos investigam a relação entre variáveis explicativas e o resultado da partida, buscando identificar fatores que influenciam as

probabilidades de vitórias com base em pontos, *games* e *sets*. A inclusão de fatores adicionais, como o desempenho passado dos jogadores, condições do jogo e outras variáveis contextuais, permite uma estimativa mais precisa da probabilidade de vitória em um ponto específico. As cadeias de Markov capturam a dependência entre pontos subsequentes, considerando a dinâmica do jogo e o histórico de resultados (BARNETT; CLARKE, 2005; MADURSKA, 2012).

Existem também estudos que utilizam técnicas de regressão para modelar diretamente a probabilidade de vitória de partidas sem considerar o ganho de pontos de forma específica. Esses modelos focam na análise de variáveis explicativas que influenciam o resultado final do jogo, como a performance geral dos jogadores, *rankings*, histórico de confrontos diretos e condições específicas da partida (por exemplo, superfície da quadra no tênis ou local do jogo no vôlei). Modelos de regressão logística, por exemplo, são amplamente utilizados para estimar a probabilidade de vitória com base em variáveis explicativas (SHUKLA; KUMAR; YADAV, 2020; WANG *et al.*, 2024). Além disso, há trabalhos que utilizam modelos de aprendizado de máquina para prever os resultados de jogos de tênis, ver por exemplo Lei, Lin e Cao (2024), Zhao, Luo e Bi (2024) e Li *et al.* (2025).

O principal modelo deste trabalho considera variáveis de Bernoulli dependentes por meio de uma cadeia de Markov. Aki e Hirano (1993) introduziram uma cadeia de Markov homogênea de dois estados, a qual foi empregada para mensurar a dependência em uma sequência de variáveis aleatórias discretas. Kolev, Minkova e Neytchev (2000) propuseram novas distribuições discretas baseadas em sequências de variáveis aleatórias binárias, e também avaliaram a dependência dentro da sequência utilizando cadeia de Markov homogênea. Dimitrov e Kolev (2002) apresentam ensaios de Bernoulli, modelando a dependência usando uma cadeia de Markov, e desenvolvem extensões com distribuições relacionadas, tais como: geométrica correlacionada, binomial correlacionada, e binomial negativa correlacionada. Da mesma forma, Souza (2019) apresenta a distribuição geométrica de ordem k correlacionada e, desenvolve modelos de regressão com os métodos de estimação clássico e bayesiano.

1.1 Objetivos e organização da tese

Neste trabalho, nosso objetivo é propor distribuições de probabilidade e correspondentes modelos de regressão para modelar o resultado de um jogo, ou seja, seus placares, e considerar a interdependência entre os *sets* por meio de uma cadeia de Markov. Inicialmente, serão propostas distribuições de probabilidade para variáveis respostas bivariadas discretas, levando em consideração a dependência entre os ensaios de Bernoulli, e também o caso particular da independência entre esses ensaios. A partir disso, desenvolvemos modelos de regressão que assumem que a variável resposta tem as distribuições de probabilidade propostas. Nesse sentido, apresentamos aspectos inferenciais, métodos de diagnóstico e aplicações dos modelos. Nossa abordagem emprega modelos de regressão para estimar a probabilidade de ocorrência de cada

um dos possíveis placares de jogos (3 a 0, 3 a 1, etc), levando em conta, especialmente, o *ranking* dos jogadores ou equipes envolvidas no jogo. Note que nossa abordagem estima a probabilidade de cada placar de jogos divididos em *sets*, enquanto os métodos propostos estimam apenas a probabilidade de vitória de cada jogador/equipe.

O trabalho está organizado da seguinte forma. No segundo capítulo, é introduzida uma distribuição bivariada para modelar um jogo esportivo dividido em *sets*, considerando independência entre os sets. São examinadas algumas propriedades dessa distribuição, e apresentamos o método de estimação por máxima verossimilhança dos parâmetros. Em seguida, desenvolvemos a modelagem de regressão assumindo que a variável resposta tem essa distribuição, obtendo o vetor escore, a matriz de informação de Fisher e discute-se a estimação pontual dos parâmetros. Também são apresentados resultados de estudos de simulação de Monte Carlo para avaliar as propriedades dos estimadores apresentados na estrutura de regressão. Por fim, o capítulo traz duas aplicações do modelo proposto e discute o ajuste por meio da análise de resíduos.

O terceiro capítulo começa definindo uma cadeia de Markov, a partir da qual é obtida a distribuição bivariada proposta. Em seguida, são detalhadas as etapas de construção e definição da distribuição para modelar dados de jogos esportivos divididos em *sets*, levando em consideração uma dependência entre os *sets* por meio de uma cadeia de Markov de dois estados. Também, são discutidos os métodos de máxima verossimilhança para estimar os parâmetros da distribuição de probabilidade. Por fim, é apresentado um breve estudo de simulação para avaliar as propriedades assintóticas dos estimadores dos parâmetros.

O quarto capítulo apresenta modelos de regressão que consideram a dependência entre os *sets*, com base na distribuição introduzida no Capítulo 3. São derivados o vetor escore e a matriz de informação de Fisher, e discute-se a estimação dos parâmetros por máxima verossimilhança. Também, são apresentadas expressões para a análise de diagnóstico do modelo, incluindo a análise de resíduos e a análise de influência global. Em seguida, realizam-se estudos de simulação de Monte Carlo para avaliar a performance dos estimadores para o modelo proposto. Por fim, o capítulo apresenta uma aplicação do modelo desenvolvido neste trabalho a dados reais de partidas de tênis, utilizando resultados de torneios do Grand Slam.

O Capítulo 5 apresenta as considerações finais que resume o que foi desenvolvido neste trabalho. Além disso, são apresentadas propostas de trabalhos futuros, com destaque para a proposição de modelos considerando outras estruturas de dependência.

CAPÍTULO

2

MODELOS DE REGRESSÃO CONSIDERANDO *SETS* INDEPENDENTES

Em jogos esportivos divididos em *sets*, o resultado final pode ser considerado uma variável aleatória bivariada. Denotamos como $(X,Y)^\top$, um vetor aleatório bidimensional discreto com características específicas. Neste contexto, X e Y representam o número de *sets* ganhos pelos jogadores A e B (ou equipes A e B), respectivamente. Para determinar o vencedor de uma partida, é necessário atingir um número predefinido de *sets*, que denotaremos como k. Este valor k é igual ao máximo entre X e Y, uma vez que um jogador vence quando alcança esse número de *sets*. Além disso, o mínimo entre X e Y é um número inteiro que varia no intervalo de 0 a $k-1$.

Inicialmente construiremos a função de probabilidade em um caso mais simples, baseada no seguinte exemplo: considere um jogo de tênis na versão melhor de cinco *sets*, o vencedor da partida é o jogador que vence três *sets*. Nesse caso, temos as seguintes definições,

$X \equiv$ número de *sets* vencido pelo jogador A;

$Y \equiv$ número de *sets* vencido pelo jogador B;

$p \equiv$ é a probabilidade de o jogador A vencer um *set*, logo $(1-p)$ é a probabilidade de o jogador B vencer um *set*.

Assumimos, na abordagem usada neste capítulo, que a probabilidade de o jogador vencer um *set* é independente de vencer outro *set*, além de ser a mesma para todos os *sets*. Nesse caso, o número máximo de *sets* para um jogador vencer o jogo é ganhar três *sets*. Então, $k = 3 = \max(X,Y)$, o $\min(X,Y) = 0,1,\ldots,k-1$ e observe que $m = 5$ é o número máximo de *sets* possíveis em uma partida.

Os possíveis resultados de uma partida, que incluem todas as combinações possíveis de vitórias em *sets* por parte de ambos jogadores ou equipes são:

$(X = 3, Y = 0) \approx$ "jogador A vence a partida em 3 *sets* consecutivos";

$(X = 3, Y = 1) \approx$ "jogador A venceu três *sets*, enquanto o jogador B venceu apenas um";

$(X = 3, Y = 2) \approx$ "jogador A venceu três *sets*, enquanto o jogador B venceu dois *sets*";

$(X = 0, Y = 3) \approx$ "jogador B que venceu três *sets* consecutivos";

$(X = 1, Y = 3) \approx$ "jogador B venceu três *sets*, enquanto o jogador A venceu apenas um";

$(X = 2, Y = 3) \approx$ "jogador B venceu três *sets*, enquanto o jogador A venceu dois *sets*".

Para exemplificar o cálculo das probabilidades, considere o interesse em calcular a probabilidade de o jogador A vencer uma partida em quatro *sets*, ou seja, o jogador A vence três *sets* enquanto o jogador B vence apenas um. A Tabela 2.1 apresenta todos os eventos possíveis para essa partida em quatro *sets*.

Tabela 2.1 – Exemplo: Jogador A vence a partida em quatro *sets*.

Eventos	1º *set*	2º *set*	3º *set*	4º *set*	probabilidade
1	A	A	B	A	$p^3(1-p)$
2	A	B	A	A	$p^3(1-p)$
2	B	A	A	A	$(1-p)p^3$

Então, a probabilidade de o jogador A vencer em quatro *sets* é $P(X = 3, Y = 1) = 3p^3(1-p)$. De forma análoga, podemos calcular todas as possíveis probabilidades para os eventos descritos anteriormente. A Tabela 2.2 apresenta essas probabilidades, de acordo com as características apresentadas. Este exemplo serve como base para a construção da função de probabilidade proposta.

Tabela 2.2 – Probabilidades para um jogo na versão melhor de cinco *sets*.

(X,Y)	$P(X = x, Y = y)$	
$(3,0)$	p^3	$= \binom{2}{2} p^3$
$(3,1)$	$3p^3(1-p)$	$= \binom{3}{2} p^3(1-p)$
$(3,2)$	$6p^3(1-p)^2$	$= \binom{4}{2} p^3(1-p)^2$
$(0,3)$	$(1-p)^3$	$= \binom{2}{2} (1-p)^3$
$(1,3)$	$3p(1-p)^3$	$= \binom{3}{2} p(1-p)^3$
$(2,3)$	$6p^2(1-p)^3$	$= \binom{4}{2} p^2(1-p)^3$

Sem perder a generalidade, assumimos em todos os capítulos deste trabalho que o jogador A é, em princípio, o favorito para vencer o jogo. A escolha do jogador favorito nas aplicações será baseada no *ranking* de jogadores ou países.

2.1 Distribuição de probabilidade bivariada proposta

Agora generalizaremos para um valor k qualquer. Seja $(X,Y)^\top$ um vetor aleatório bivariado discreto, representando os placares de um resultado de um jogo esportivo dividido em *sets*,

considerando independência entre os *sets*. Então, a função de probabilidade conjunta é definida por

$$P(X=x,Y=y)=\begin{cases}\binom{x+y-1}{k-1}p^x(1-p)^y, & \text{se } \max(x,y)=k \text{ e } \min(x,y)=0,\ldots,k-1\\ 0, & \text{caso contrário}\end{cases} \tag{2.1}$$

em que $x,y=0,1,\ldots,k$ com $k=1,2,3,...$ e $0<p<1$.

A partir da equação (2.1), podemos definir as distribuições de probabilidade marginais de X e Y. Elas são dadas por

$$P(X=x)=\sum_{y=0}^{k}P(X=x,Y=y); \tag{2.2}$$

$$P(Y=y)=\sum_{x=0}^{k}P(X=x,Y=y). \tag{2.3}$$

A partir das equações (2.2) e (2.3), pode-se notar que as probabilidades de vitória de cada um dos jogadores é dada por

$$\begin{aligned}P(X=k) &= \sum_{i=0}^{k-1}P(X=k,Y=i)\\ P(Y=k) &= \sum_{i=0}^{k-1}P(X=i,Y=k).\end{aligned}$$

Também podemos obter os momentos das distribuições de probabilidades marginais. Temos que a esperança e variância dessas variáveis são dados por

$$\mu_X = E(X)=\sum_x xP(X=x)=\sum_{i=0}^{k-1}iP(X=i,Y=k)+k\sum_{j=0}^{k-1}P(X=k,Y=j), \tag{2.4}$$

$$\mu_Y = E(Y)=\sum_y yP(Y=y)=\sum_{i=0}^{k-1}iP(X=k,Y=i)+k\sum_{j=0}^{k-1}P(X=j,Y=k), \tag{2.5}$$

$$\begin{aligned}E(X^2) &= \sum_x x^2P(X=x)=\sum_{i=0}^{k-1}i^2P(X=i,Y=k)+k^2\sum_{j=0}^{k-1}P(X=k,Y=j),\\ E(Y^2) &= \sum_y y^2P(Y=y)=\sum_{i=0}^{k-1}i^2P(X=k,Y=i)+k^2\sum_{j=0}^{k-1}P(X=j,Y=k).\end{aligned}$$

Assim, a variância é obtida por $VAR(X)=E(X^2)-[E(X)]^2$. A demonstração do cálculo das esperanças para um caso particular está apresentada no Anexo A.

2.1.1 Estimador de máxima verossimilhança

Considerando $(x_1,y_1),\ldots,(x_n,y_n)$ uma amostra aleatória de tamanho n, obtida das variáveis aleatórias bivariadas $(X_1,Y_1),\ldots,(X_n,Y_n)$. A função de verossimilhança para p é dada por

$$\begin{aligned} L(p|(\mathbf{x},\mathbf{y})) &= \prod_{i=1}^{n} P(X=x_i, Y=y_i) \\ &= \left\{\prod_{i=1}^{n}\binom{x_i+y_i-1}{k-1}\right\} p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n} y_i}. \end{aligned}$$

Aplicando o logaritmo, obtemos o logaritmo da função de verossimilhança, dado por

$$\ell(p|(\mathbf{x},\mathbf{y})) = \sum_{i=1}^{n}\log\binom{x_i+y_i-1}{k-1} + \sum_{i=1}^{n} x_i\log(p) + \sum_{i=1}^{n} y_i\log(1-p). \tag{2.6}$$

O estimador de máxima verossimilhança (EMV) para o parâmetro p é obtido a partir da maximização do logaritmo da função de verossimilhança. Assim, derivando (2.6) em relação a p, encontramos a função escore dada por

$$\frac{\partial}{\partial p}\ell(p|(\mathbf{x},\mathbf{y})) = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n} y_i}{(1-p)}.$$

A partir dessa função, igualando a zero, obtemos o seguinte estimador de máxima verossimilhança:

$$\frac{\sum_{i=1}^{n} x_i}{\widehat{p}} - \frac{\sum_{i=1}^{n} y_i}{(1-\widehat{p})} = 0,$$

então,

$$\widehat{p} = \frac{\overline{x}}{\overline{x}+\overline{y}}.$$

O estimador $\widehat{p}$ representa a proporção amostral de *sets* vencidas pelos jogadores ou equipes A. Esse estimador era esperado, pois, fixado um *set* qualquer do jogo, p representa a probabilidade do jogador ou equipe A vencer esse *set*.

2.1.2 Informação de Fisher

A derivada de segunda ordem do logaritmo da função de verossimilhança com relação ao parâmetro p é dada por

$$\frac{\partial^2}{\partial p^2}\ell(p|(\mathbf{x},\mathbf{y})) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{\sum_{i=1}^n y_i}{(1-p)^2}.$$

Desta forma, a informação de Fisher para o parâmetro p, é dada por

$$I_F(p) = -E\left(\frac{\partial^2}{\partial p^2}\ell(p|(\mathbf{x},\mathbf{y}))\right) = \frac{\sum_{i=1}^n E(X_i)}{p^2} + \frac{\sum_{i=1}^n E(Y_i)}{(1-p)^2},$$

em que $E(X)$ e $E(Y)$ são apresentados em (2.4) e (2.5), respectivamente. Como exemplo, fixamos $k=4$, logo a informação de Fisher é dada por

$$I_F(p) = n\left[\frac{\mu_X}{p^2} + \frac{\mu_Y}{(1-p)^2}\right], \tag{2.7}$$

em que

$$\begin{aligned}\mu_X = {} & 4p(1-p)^4 + 20p^2(1-p)^4 + 60p^3(1-p)^4 + \\ & 4p^4 + 16p^4(1-p) + 40p^4(1-p)^2 + 80p^4(1-p)^3,\end{aligned} \tag{2.8}$$

e

$$\begin{aligned}\mu_Y = {} & 4p^4(1-p) + 20p^4(1-p)^2 + 60p^4(1-p)^3 + \\ & 4(1-p)^4 + 16p(1-p)^4 + 40p^2(1-p)^4 + 80p^3(1-p)^4.\end{aligned} \tag{2.9}$$

As demostrações das expressões (2.8) e (2.9) são apresentadas no Anexo A. A partir dessas expressões, podemos obter a distribuição assintótica do estimador de máxima verossimilhança para p. Bolfarine e Sandoval (2001) apresentam a distribuição para grandes amostras, dada por $\sqrt{n}(\widehat{p}-p) \xrightarrow{D} N\left(0, I_F(p)^{-1}\right)$, em que $\widehat{p}$ é o estimador de máxima verossimilhança do parâmetro p, e $I_F(p)$ é a informação de Fisher dada em (2.7).

2.2 Modelos de regressão

Nesta seção, introduzimos um modelo de regressão que assume que a variável resposta tem a distribuição de probabilidade conjunta proposta em (2.1). Obtêm-se o vetor escore e a matriz de informação de Fisher e discute-se a estimação dos parâmetros pelo método de máxima verossimilhança.

Suponha $(X_1,Y_1),\ldots,(X_n,Y_n)$ variáveis aleatórias independentes em que (X_i,Y_i) tem distribuição dada por (2.1). Assumimos que $(x_1,y_1),\ldots,(x_n,y_n)$ são observações de $(X_1,Y_1),\ldots,(X_n,Y_n)$, respectivamente. Portanto, modelamos p_i em função de covariáveis utilizando o seguinte componente sistemático

$$g(p_i) = \sum_{j=0}^{J} \beta_j z_{ij} = \mathbf{z}_i^\top \boldsymbol{\beta} = \eta_i, \tag{2.10}$$

em que β_j são parâmetros desconhecidos, $j = 0, \ldots, J$, e $z_{i0} = 1$ para todo i, com β_0 representando o intercepto do modelo, $z_{i1}, z_{i2}, \ldots, z_{iJ}$ são observações das covariáveis Z_{ij}, e $g(.)$ é a função de ligação, com g sendo uma função estritamente monótona e duplamente diferenciável. Portanto, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_J)^\top$ é o vetor de coeficientes da regressão, com $\boldsymbol{\beta} \in \mathbb{R}^r$, $r = J+1$, tal que $r+1 < n$, e η_i é o preditor linear e $\mathbf{z}_i = (1, z_{i1}, \ldots, z_{iJ})^\top$ o vetor de variáveis regressoras de ordem r.

Sendo $g : (0,1) \to \mathbb{R}$ a função de ligação, é comum escolher diferentes funções para essa finalidade. Entre as funções de ligação mais conhecidas estão a logito, frequentemente utilizada por gerar parâmetros interpretáveis, a probito, a complemento log-log e a log-log. McCullagh e Nelder (1989) definem essas funções de ligação da seguinte forma: logito, $g(p_i) = \log\left(\frac{p_i}{(1-p_i)}\right)$, probito, $g(p_i) = \Phi(p_i)^{-1}$, em que $\Phi(.)$ é a função distribuição acumulada da normal padrão, complemento log-log, $g(p_i) = \log[-\log(1-p_i)]$, e log-log, $g(p_i) = -\log[-\log(p_i)]$. Outras diversas funções de ligação já foram propostas para parâmetros com espaço paramétrico no intervalo $(0,1)$, tais como função potência, potência logito, potência probito, entre outras. Para mais detalhes ver Bazán *et al.* (2017), Lemonte e Bazán (2018).

2.2.1 Estimação dos parâmetros

Os estimadores do modelo podem ser obtidos pelo método de máxima verossimilhança. As estimativas dos parâmetros são tais que a probabilidade conjunta dos valores observados é máxima. O logaritmo da função de verossimilhança de $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_J)^\top$ dada uma amostra observada $(\mathbf{x}, \mathbf{y}) = (x_1, y_1), \ldots, (x_n, y_n)$, pode ser escrito como

$$\ell(\boldsymbol{\beta}|(\mathbf{x},\mathbf{y})) = \sum_{i=1}^{n} \left\{ \log \binom{x_i + y_i - 1}{k-1} + x_i \log(p_i) + y_i \log(1-p_i) \right\}. \tag{2.11}$$

A fim de obter a função escore associada a uma amostra $(\mathbf{x}, \mathbf{y})$, deriva-se (2.11) em relação a cada componente de $\boldsymbol{\beta}$, ou seja,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta}|(\mathbf{x},\mathbf{y}))}{\partial \beta_j} &= \sum_{i=1}^{n} \left\{ \frac{x_i}{p_i} \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} - \frac{y_i}{(1-p_i)} \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \right\} \\ &= \sum_{i=1}^{n} \left\{ \left[\frac{x_i}{p_i} - \frac{y_i}{(1-p_i)} \right] z_{ij} \frac{\partial p_i}{\partial \eta_i} \right\} \\ &= \sum_{i=1}^{n} h_i f_i z_{ij}, \end{aligned}$$

em que $h_i = x_i/p_i - y_i/(1-p_i)$; $f_i = \partial p_i/\partial \eta_i$. Então, podemos escrever a função escore na sua forma matricial,

$$U(\boldsymbol{\beta}) = Z^\top H F,$$

em que Z é a matriz modelo com dimensão $n \times r$ e possui linhas $\mathbf{z}_i^\top$, para $i = 1, \ldots, n$; $H = \text{diag}\{h_1, \ldots, h_n\}$ é uma matriz diagonal, e $F = (f_1, \ldots, f_n)^\top$ é um vetor de tamanho n.

A matriz de informação de Fisher $J(\boldsymbol{\beta})$ é obtida, calculando-se a derivada de segunda ordem, cujos elementos são dados por

$$\frac{\partial^2 \ell(\boldsymbol{\beta}|(\mathbf{x},\mathbf{y}))}{\partial\beta_j\partial\beta_k} = \sum_{i=1}^{n}\left\{\left[\left[-\frac{x_i}{p_i^2}-\frac{y_i}{(1-p_i)^2}\right]\left(\frac{\partial p_i}{\partial\eta_i}\right)^2+\left[\frac{x_i}{p_i}-\frac{y_i}{(1-p_i)}\right]\left(\frac{\partial^2 p_i}{\partial\eta_i^2}\right)\right] z_{ij}z_{ik}\right\},$$

cujos valores esperados são obtidos por

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta}|(\mathbf{x},\mathbf{y}))}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n}\left\{\left[\left[-\frac{E(X_i)}{p_i^2}-\frac{E(Y_i)}{(1-p_i)^2}\right]\left(\frac{\partial p_i}{\partial\eta_i}\right)^2+\left[\frac{E(X_i)}{p_i}-\frac{E(Y_i)}{(1-p_i)}\right]\left(\frac{\partial^2 p_i}{\partial\eta_i^2}\right)\right] z_{ij}z_{ik}\right\}.$$

Logo, podemos expressar a informação de Fisher por

$$\begin{aligned} J_{\beta_j\beta_k} &= -E\left(\frac{\partial^2 \ell(\boldsymbol{\beta}|(\mathbf{x},\mathbf{y}))}{\partial\beta_j\partial\beta_k}\right) \\ &= \sum_{i=1}^{n}\left\{\left[\left[\frac{E(X_i)}{p_i^2}+\frac{E(Y_i)}{(1-p_i)^2}\right]\left(\frac{\partial p_i}{\partial\eta_i}\right)^2-\left[\frac{E(X_i)}{p_i}-\frac{E(Y_i)}{(1-p_i)}\right]\left(\frac{\partial^2 p_i}{\partial\eta_i^2}\right)\right] z_{ij}z_{ik}\right\} \\ &= \sum_{i=1}^{n}\left(b_i f_i^2 - a_i d_i\right) z_{ij}z_{ik} \\ &= \sum_{i=1}^{n} v_i z_{ij}z_{ik}, \end{aligned}$$

em que $v_i = b_i f_i^2 - a_i d_i$; $b_i = E(X_i)/p_i^2 + E(Y_i)/(1-p_i)^2$; $a_i = E(X_i)/p_i - E(Y_i)/(1-p_i)$; $d_i = \partial^2 p_i/\partial\eta_i^2$. As médias $E(X_i)$ e $E(Y_i)$ foram definidas em (2.4) e (2.5), respectivamente.

Desta forma, podemos escrever a informação de Fisher de $\boldsymbol{\beta}$ na forma matricial,

$$J(\boldsymbol{\beta}) = Z^\top V Z, \tag{2.12}$$

sendo $V = \text{diag}\{v_1,\ldots,v_n\}$ a matriz diagonal que traz informação da distribuição e da função de ligação usada. A Tabela 2.3 apresenta as derivadas de primeira e segunda ordem referente às principais funções de ligação.

Devido à falta de solução analítica explícita para o sistema de equações, os estimadores de máxima verossimilhança dos parâmetros do modelo, denotados por $\widehat{\boldsymbol{\beta}}$, são obtidos por métodos numéricos. Todos os estudos de simulação pelo método de Monte Carlo, bem como as aplicações, utilizaram o método de aproximação numérica quase-Newton BFGS (YANG; SMALL, 2013).

Tabela 2.3 – Derivadas de primeira e segunda ordem da função de ligação em relação aos parâmetros $\boldsymbol{\beta}$.

Função de ligação	$p_i = g^{-1}(\eta_i)$	$f_i = \partial p_i/\partial\eta_i$	$d_i = \partial^2 p_i/\partial\eta_i^2$
Logito	$\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$	$\exp(\eta_i)(1+\exp(\eta_i))^{-2}$	$\exp(\eta_i)[1-\exp(\eta_i)](1+\exp(\eta_i))^{-3}$
C.log-log	$1-\exp(-\exp(\eta_i))$	$\exp(\eta_i-\exp(\eta_i))$	$[\exp(\eta_i-\exp(\eta_i))-\exp(2\eta_i-\exp(\eta_i))]$
Log-log	$\exp(-\exp(-\eta_i))$	$\exp(-\eta_i-\exp(-\eta_i))$	$-[\exp(-\eta_i-\exp(-\eta_i))-\exp(-2\eta_i-\exp(-\eta_i))]$
Probito	$\Phi(\eta_i)$	$\phi(\eta_i)$	$-\eta_i\phi(\eta_i)$

Em que $\Phi(\cdot)$ e $\phi(\cdot)$ são as funções de distribuição acumulada e densidade da normal padrão, respectivamente.

2.3 Intervalos de confiança e testes de hipóteses assintóticos

Para grandes amostras e a função de distribuição conjunta proposta em (2.1) satisfazendo as condições de regularidades usuais (MILLAR, 2011, capítulo 12), podemos fazer uso das propriedades assintóticas dos estimadores. Sendo $\widehat{\boldsymbol{\beta}}$ e $J(\widehat{\boldsymbol{\beta}})$ estimadores de máxima verossimilhança para $\boldsymbol{\beta}$ e $J(\boldsymbol{\beta})$, respectivamente, temos que

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{D} N_r\left(\mathbf{0}, J(\widehat{\boldsymbol{\beta}})^{-1}\right),$$

em que $J(\widehat{\boldsymbol{\beta}})$ é a informação de Fisher, conforme definido em (2.12) e avaliada em $\widehat{\boldsymbol{\beta}}$. Portanto, para grandes amostras, pela normalidade assintótica do estimador de máxima verossimilhança $\widehat{\beta}_j$, o intervalo de confiança para cada um dos parâmetros com coeficiente de confiança de $100(1-\alpha)\%$ é dado por

$$IC(\beta_j; 1-\alpha) = \widehat{\beta}_j \pm z_{\alpha/2}\left(J^{\widehat{\beta}_j\widehat{\beta}_j}\right)^{1/2},$$

em que $z_{\alpha/2}$ é o quantil $1-\alpha/2$ da distribuição normal padrão e, $\left(J^{\widehat{\beta}_j\widehat{\beta}_j}\right)$ é o termo de posição j da diagonal de $J(\widehat{\boldsymbol{\beta}})^{-1}$, que corresponde à variância estimada do estimador de máxima verossimilhança do parâmetro β_j, com $j = 1,\ldots,r$.

Para testar hipóteses a respeito dos parâmetros do modelo proposto, ou seja, testar a hipótese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ (hipótese nula), em que $\boldsymbol{\beta}_0$ é um vetor r-dimensional de valores fixados, contra a hipótese alternativa $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\beta}_0$. Portanto, usando o teste de Wald a estatística é dada por

$$W = \left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)^\top J(\widehat{\boldsymbol{\beta}})\left(\widehat{\beta}-\boldsymbol{\beta}_0\right).$$

Esta estatística de teste, sob a hipótese nula, possui distribuição assintótica qui-quadrado com r graus de liberdade, denotada por $\chi^2_{(r)}$, em que r é a quantidade de parâmetros, que em nosso modelo é $r = J+1$.

Quando há interesse em avaliar se um único parâmetro específico do modelo é significativamente diferente de um valor, pode-se também utilizar a estatística do teste de Wald. Neste caso, a estatística de Wald simplifica-se bastante sendo calculada como o quadrado da diferença entre a estimativa do parâmetro e o valor testado β_j^*, dividido pela estimativa da variância do estimador. Matematicamente, se $\widehat{\beta}_j$ é a estimativa do parâmetro, a estatística de Wald é dada por:

$$W_j = \frac{(\widehat{\beta}_j-\beta_j^*)^2}{\left(J^{\widehat{\beta}_j\widehat{\beta}_j}\right)}.$$

Sob a hipótese nula, essa estatística segue uma distribuição qui-quadrado com um grau de liberdade. Rejeitamos a hipótese nula se o valor calculado de W_j exceder o quantil da distribuição qui-quadrado correspondente a $1-\alpha$.

Além do teste de Wald, também podem ser usados os testes da razão de verossimilhança, escore e gradiente (LEHMANN; ROMANO; CASELLA, 1986; LEMONTE, 2016). O teste da razão de verossimilhança é particularmente conveniente quando se deseja testar mais de um parâmetro simultaneamente. Quando estamos testando um número b de parâmetros, a estatística do teste da razão de verossimilhança é calculada como $RV = -2\left(\log L(\widehat{\boldsymbol{\beta}}_0) - \log L(\widehat{\boldsymbol{\beta}})\right)$, em que $L(\widehat{\boldsymbol{\beta}}_0)$ é a função de verossimilhança sob a hipótese nula e $L(\widehat{\boldsymbol{\beta}})$ é a função de verossimilhança sob a hipótese alternativa. Esta estatística, sob a hipótese nula, segue uma distribuição qui-quadrado com b graus de liberdade.

2.4 Análise de diagnóstico

Os resultados apresentados nas seções anteriores permitem o ajuste de um modelo de regressão para dados reais de jogos esportivos divididos em *sets*, além da realização da análise inferencial do modelo obtido. No entanto, as conclusões obtidas serão validadas somente se o ajuste do modelo for adequado. Nesta seção, discutimos a análise de diagnóstico e, principalmente, os gráficos de resíduos que podem auxiliar na avaliação da adequabilidade e robustez dos resultados inferenciais do modelo. Os resíduos representam desvios entre os valores observados e os valores ajustados, podendo indicar um modelo inadequado, conforme discutido por Paula (2025).

O resíduo quantílico aleatorizado foi proposto por Dunn e Smyth (1996) para avaliar a qualidade do ajuste de modelos. São considerados uma boa opção de resíduos quando a variável resposta é discreta e assume poucos valores diferentes. Esses resíduos são obtidos a partir do uso de duas funções de distribuição acumuladas, a da normal padrão e da suposta para a variável resposta.

Aqui, vamos fazer uso da transformação $w_i = x_i - y_i$, que é uma função bijetora, ou seja, que garante que para cada valor de w_i, seja possível obter o par de valores originais (x_i, y_i) e vice-versa. Seja, $F(w_i; \widehat{p})$ a função de distribuição acumulada de W_i, sendo esta uma função não contínua. Os resíduos quantílicos aleatorizados, denotados por r_{q_i} são definidos por

$$r_{q_i} = \Phi^{-1}(u_i),$$

em que Φ é a função de distribuição acumulada da normal padrão, e u_i é uma variável aleatória com distribuição uniforme no intervalo $(a_i, b_i]$, sendo a_i e b_i obtidos por

$$a_i = F(w_i - 1; \widehat{p}) \qquad \text{e} \qquad b_i = F(w_i; \widehat{p}).$$

Os resíduos quantílicos aleatorizados possuem distribuição assintótica normal padrão se o modelo for o correto, e seus parâmetros forem estimados de forma consistente. Portanto, podemos usar o gráfico de probabilidade normal (*QQplot*), que mostra os resíduos em função dos quantis da distribuição normal padrão, para identificar possíveis discrepâncias entre os valores

observados e esperados para os resíduos. Como os resíduos não são independentes, é difícil na prática saber se desvios entre os valores observados e esperados para os resíduos indicam ou não uma falta de ajuste. Assim, é interessante associar o gráfico de probabilidade normal padrão com um envelope simulado (ATKINSON, 1985).

2.5 Estudos de simulação

Nesta seção, é apresentado um breve estudo de simulação com o objetivo de avaliar a performance do estimador de máxima verossimilhança no modelo de regressão definido em (2.11). Os algoritmos para geração da distribuição bivariada proposta, o ajuste do modelo de regressão e as métricas usadas na avaliação da estimação foram desenvolvidos no *software* R (R Core Team, 2023). As estimativas de máxima verossimilhança dos parâmetros do modelo foram calculadas e seu desempenho avaliado por meio do viés e da raiz do erro quadrático médio (REQM). Para um parâmetro β_j qualquer, essas são definidas por $\widehat{B(\beta_j)} = \widehat{E(\beta_j)} - \beta_j$, com $\widehat{E(\beta_j)} = S^{-1}\sum_{i=1}^{S}\widehat{\beta_{ij}}$, em que $\widehat{\beta_{ij}}$ é a estimativa de máxima verossimilhança de β_j na i-ésima réplica de Monte Carlo, REQM $= \sqrt{S^{-1}\sum_{i=1}^{S}\left(\widehat{\beta_{ij}} - \beta_j\right)^2}$, em que S é o número de réplicas de Monte Carlo.

O conjunto de dados com distribuição bivariada proposta (2.1) foram simulados considerando $k = 4$, ou seja, simulando-se um jogo na versão melhor de sete *sets*, que ganha a partida um jogador (ou equipe) que vencer quatro *sets*. Foram avaliados quatro diferentes cenários, obtidos de quatro diferente valores para os vetores de parâmetros. Duas covariáveis foram consideradas, com distribuições $Z_{i1} \sim U(0,1)$ e $Z_{i2} \sim U(0,1)$, $i = 1,\ldots,n$, e estas foram mantidas constantes em todas as S réplicas de Monte Carlo.

O estudo foi realizado por meio de simulação de Monte Carlo com $S = 5000$ réplicas, considerando os seguintes tamanhos amostrais, $n = 20, 40, 60, 100, 500, 1000$. A função de ligação logito foi escolhida para o modelo de regressão, por ser particularmente conveniente quando o parâmetro possui espaço paramétrico no intervalo $(0,1)$, uma vez que proporciona estimativas interpretáveis dos parâmetros. Os cenários do estudo de simulação foram definidos de modo a avaliar as características do modelo com base nos valores extremos das variáveis regressoras estabelecidas. As simulações foram planejadas variando-se as diferenças entre as probabilidades de vitória em um *set* para os melhores e piores jogadores, bem como a intensidade do efeito das variáveis explicativas sobre a variável resposta. Cabe destacar que, quanto maior o valor absoluto dos parâmetros $\boldsymbol{\beta}$(exceto β_0) mais intenso é o efeito das covariáveis sobre a variável resposta. No primeiro cenário, os coeficientes de regressão foram definidos como $\beta_0 = 0,2$, $\beta_1 = 0,5$ e $\beta_2 = -0,5$ indicando que os melhores jogadores, em média, apresentam desempenho ligeiramente superior aos piores jogadores, e que as variáveis regressoras exercem um efeito fraco sobre a variável resposta. No segundo cenário, os coeficientes foram $\beta_0 = 0,2$, $\beta_1 = 1,5$ e $\beta_2 = -1,5$, mostrando que os melhores jogadores são, em média, ligeiramente

melhores que os piores jogadores, e que as variáveis regressoras têm efeito forte sobre a variável resposta. No terceiro cenário, os coeficientes foram $\beta_0 = 1,0$, $\beta_1 = 0,5$ e $\beta_2 = -0,5$, sugerindo que os melhores jogadores, em média, são consideravelmente melhores que os piores jogadores, e que as covariáveis apresentam efeito fraco sobre a variável resposta. Por fim, no quarto cenário, os coeficientes da regressão foram $\beta_0 = 1,0$, $\beta_1 = 1,5$ e $\beta_2 = -1,5$, representando que os melhores jogadores são, em média, consideravelmente melhores que os piores jogadores, e que as covariáveis tem efeito forte sobre a variável resposta.

Na Tabela 2.4, encontram-se os resultados obtidos das simulações para o modelo de regressão proposto (2.11) para a distribuição bivariada no caso independente (2.1). Observa-se que as estimativas dos vieses são próximas de zero e assim como as estimativas do REQM, diminuem à medida que o tamanho amostral aumenta, conforme esperado.

Os dois primeiros cenários apresentaram desempenho bastante semelhante em relação à performance dos estimadores. Nota-se que os estimadores são ligeiramente viciados para pequenas amostras, e parecem ser não viciadas ou com viés pequeno para amostras grandes. Além disso, a REQM das estimativas nos dois primeiros cenários foi menor em comparação aos dois últimos. Nestes últimos cenários, nota-se que, para amostras pequenas, o viés dos estimadores é superior ao observado anteriormente; contudo, ambos os cenários também apresentam viés reduzido e REQM pequeno para tamanhos amostrais elevados. Esses resultados sugerem que o REQM dos estimadores do modelo proposto tende a ser menor quando a diferença entre as probabilidades de vitória dos melhores e piores jogadores é pequena.

Tabela 2.4 – Estimativas do viés e REQM (em parênteses) dos estimadores dos parâmetros nos dados simulados do modelo considerando *sets* independentes.

n	$\widehat{\beta}_0\,(0,2)$	$\widehat{\beta}_1\,(0,5)$	$\widehat{\beta}_2\,(-0,5)$
20	0,0246 (0,7314)	0,0248 (0,9894)	−0,0500 (0,6791)
40	0,0077 (0,4274)	0,0155 (0,5330)	−0,0178 (0,5011)
60	0,0072 (0,3695)	0,0113 (0,4183)	−0,0158 (0,4213)
100	−0,0006 (0,2774)	0,0139 (0,3274)	−0,0075 (0,3033)
500	−0,0028 (0,0993)	0,0039 (0,1343)	0,0011 (0,1343)
1000	−0,0023 (0,0692)	0,0018 (0,0922)	0,0020 (0,0902)
n	$\widehat{\beta}_0\,(0,2)$	$\widehat{\beta}_1\,(1,5)$	$\widehat{\beta}_2\,(-1,5)$
20	0,0160 (0,7585)	0,1006 (1,0578)	−0,1080 (0,7436)
40	0,0015 (0,4425)	0,0485 (0,5728)	−0,0449 (0,5356)
60	0,0026 (0,3778)	0,0361 (0,4508)	−0,0290 (0,4485)
100	0,0106 (0,2861)	0,0118 (0,3520)	−0,0302 (0,3344)
500	−0,0009 (0,1000)	0,0044 (0,1437)	−0,0024 (0,1436)
1000	−0,0009 (0,0732)	0,0010 (0.1011)	0,0002 (0,0987)
n	$\widehat{\beta}_0\,(1,0)$	$\widehat{\beta}_1\,(0,5)$	$\widehat{\beta}_2\,(-0,5)$
20	0,0627 (0,8697)	0,0377 (1,1429)	−0,0457 (0,8241)
40	0,0403 (0,5000)	0,0054 (0,6256)	−0,0278 (0,6077)
60	0,0265 (0,4275)	−0,0011 (0,4890)	0,0163 (0,4983)
100	0,0158 (0,3226)	0,0028 (0,3814)	−0,0105 (0,3545)
500	0,0011 (0,1162)	−0,0010 (0,1560)	0,0025 (0,1595)
1000	−0,0005 (0,0827)	0,0001 (0,1095)	0,0021 (0,1082)
n	$\widehat{\beta}_0\,(1,0)$	$\widehat{\beta}_1\,(1,5)$	$\widehat{\beta}_2\,(-1,5)$
20	0,0712 (0,9029)	0,0894 (1,2402)	−0,1046 (0,8713)
40	0,0396 (0,5313)	0,0541 (0,6355)	−0,0742 (0,6526)
60	0,0237 (0,4295)	0,0228 (0,4879)	−0,0293 (0,5339)
100	0,0103 (0,3279)	0,0249 (0,4054)	−0,0231 (0,3744)
500	−0,0005 (0,1175)	0,0075 (0,1602)	−0,0019 (0,1672)
1000	0,0018 (0,0829)	−0,0007 (0,1135)	−0,0005 (0,1118)

2.6 Aplicações

Nesta seção, com o objetivo de ilustrar o modelo de regressão proposto sob a suposição de independência entre os *sets*, são apresentadas duas aplicações: a primeira com dados de tênis e a segunda com dados de tênis de mesa.

2.6.1 Dados de tênis: Copa Davis

A Copa Davis é uma das competições mais antigas e prestigiadas do tênis mundial, que reúne os melhores tenistas masculinos do mundo. Fundada em 1900, a Copa Davis possui um extenso histórico que contempla mais de um século de competições entre as nações. Os torneios

são disputados entre os países e cada confronto envolve até quatro disputas individuais e uma disputa de duplas. Os confrontos individuais são disputados em melhor de três *sets*, ou seja, ganha a partida o jogador que fizer dois *sets*. As partidas de duplas são disputadas em melhor de três *sets*. Neste torneio, os confrontos entre os países são disputados no melhor de cinco jogos (ou partidas), em que o país que vence três jogos ganha a disputa da rodada no torneio. Portanto, neste caso, temos $k = 3$.

O conjunto de dados foi obtido a partir dos resultados das eliminatórias da Copa Davis, realizadas em casa e fora, nos dias 1 a 4 de fevereiro de 2024. As informações para a construção da base de dados foram coletadas do site oficial da Copa Davis, acessível por meio do seguinte site: <www.daviscup.com/en/draws-results/qualifiers/2024.aspx>. As variáveis selecionadas das eliminatórias incluem o resultado das partidas, o *ranking* mundial dos países e o *ranking* dos jogadores.

O *ranking* é uma medida que indica a posição relativa de um jogador (ou equipe) em relação a outros jogadores (ou equipes) em termos do desempenho histórico. Essa medida é frequentemente utilizada para determinar a classificação dos jogadores em competições, bem como para comparar o desempenho ao longo do tempo.

Para obter as variáveis explicativas referentes ao *ranking* dos jogadores, faremos uso do *ranking* dos jogadores antes do torneio considerado nos dados. Isso ocorre porque o *ranking* fornecido no site da Copa Davis é atual e não corresponde à data do torneio em questão. Assim, obtivemos o *ranking* dos jogadores do mês de janeiro de 2024, disponível no site: <https://www.atptour.com/en/rankings/singles>. Da mesma forma, obtivemos o *ranking* nacional dos países antes do evento, acessível através do site: <https://archive.is/hVdCA>. A Tabela 2.5 apresenta os dados referentes à Copa Davis, construídos para aplicação do modelo proposto.

A Copa Davis é disputada em um formato de confronto com características específicas. Cada país participante forma uma equipe que pode incluir até cinco jogadores, selecionados com base no *ranking* e desempenho recentes. Os resultados dos placares de uma rodada são obtidos por meio de confrontos de partidas individuais e duplas. Por exemplo, a primeira linha de informação da Tabela 2.5 mostra o placar de 1 a 3 (placar A e placar B). Nessa primeira rodada, foram disputados quatro jogos (ou quatro partidas), sendo três delas partidas individuais (jogador A vs. jogador B) e uma partida de duplas (dupla A vs. dupla B).

Essa característica de jogo disputado por tenistas diferentes auxilia na suposição de independência entre os jogos, ou seja, é razoável assumir que a probabilidade de um jogador ganhar um jogo é independente da probabilidade de um jogador do mesmo país ganhar outro jogo.

Tabela 2.5 – Observações das eliminatórias da Copa Davis 2024.

ID	RS_BEST_A	RS_2BEST_A	RD_BEST_A	R_NAT_A	PLACAR_A	PLACAR_B	R_NAT_B	RS_BEST_B	RS_2BEST_B	RD_BEST_B
1	224	700	140	18	1	3	2	132	486	553
2	41	53	66	5	0	3	23	122	142	255
3	131	192	23	17	3	1	6	125	1044	4
4	24	60	15	8	3	1	22	57	82	512
5	29	63	7	9	3	2	21	175	199	257
6	31	127	31	10	3	0	31	415	465	205
7	32	33	3	11	3	0	35	342	498	947
8	55	166	31	12	3	0	26	47	245	49
9	17	68	10	13	3	0	32	230	238	271
10	22	25	13	19	3	2	14	278	338	48
11	119	241	57	25	3	1	15	160	364	67
12	87	438	181	34	2	3	16	20	54	745

As informações relacionadas aos dados são:

- RS_BEST_A : é o *ranking* do jogador do país A mais bem posicionado no *ranking*;
- RS_2BEST_A: é o *ranking* do segundo jogador do país A mais bem posicionado no *ranking*;
- RD_BEST_A: é o *ranking* do jogador mais bem posicionado no *ranking* da dupla do país A;
- R_NAT_A: É o *ranking* nacional do país A;
- PLACAR_A: é o número de jogos vencidos pelo país A;
- PLACAR_B: é o número de jogos vencidos pelo país B;
- R_NAT_B: É o *ranking* nacional do país B;
- RS_2BEST_B: é o *ranking* do segundo jogador do país B mais bem posicionado no *ranking*;
- RS_BEST_B : é o *ranking* do jogador do país B mais bem posicionado no *ranking*;
- RD_BEST_B: é o *ranking* do jogador mais bem posicionado no *ranking* da dupla do país B.

Para análise e determinação das variáveis explicativas, baseadas na correlação com a variável resposta, são necessárias transformações nas variáveis. Por exemplo, inicialmente definimos a variável "MR", que representa a média dos *rankings* dos jogadores individuais e das duplas para seus respectivos times. Definimos as seguintes variáveis:

- $w_i = x_i - y_i$: diferença entre os placares (PLACAR_A-PLACAR_B);

- MR_A = $\frac{\text{RS\_BEST\_A+RS\_2BEST\_A+RD\_BEST\_A}}{3}$: é a média dos *ranking* dos jogadores do país A;
- MR_B = $\frac{\text{RS\_BEST\_B+RS\_2BEST\_B+RD\_BEST\_B}}{3}$: é a média dos *ranking* dos jogadores do país B;
- z_{i1} = MR_B-MR_A : diferença entre a média dos *rankings*;
- z_{i2} = R_NAT_B-R_NAT_A : diferença entre o *ranking* nacional dos países;
- z_{i3} = MR_B/MR_A : razão entre a média dos *rankings*;
- z_{i4} = R_NAT_B/R_NAT_A : razão entre o *ranking* nacional dos países.

Os dados apresentados na Tabela 2.5 foram reorganizados de modo que o país (ou equipe) com melhor média de *ranking* fique selecionado ao longo do eixo x na análise bivariada (x,y). Essa reorganização facilita a análise e a interpretação dos dados. A Tabela 2.6 apresenta a distribuição de frequência para a variável resposta, representando os resultados dos placares das eliminatórias da Copa Davis de 2024, observamos que os placares "três a zero" e "três a um" foram os mais frequentes na amostra.

Tabela 2.6 – Distribuição de frequências para os resultados dos placares da Copa Davis.

(x,y)	Freq. absoluta	Freq. relativa
(3,0)	4	0,34
(3,1)	3	0,25
(3,2)	2	0,17
(0,3)	1	0,08
(1,3)	1	0,08
(2,3)	1	0,08
Total	12	1

Uma maneira de avaliar possíveis escolhas de variáveis explicativas é observar a correlação entre essas variáveis com a variável resposta. A Tabela 2.7 apresenta a matriz de correlação das possíveis covariáveis definidas anteriormente. Nota-se que a variável explicativa z_{i1} apresenta a maior correlação com a resposta w_i, seguida por z_{i2}. Conforme esperado, observamos que existe uma relação entre as covariáveis, com alta correlação entre z_{i1} e z_{i3}, entre z_{i2} e z_{i4}, já que cada um desses pares de covariáveis são funções dos mesmos *rankings*. Também há uma moderada correlação entre z_{i1} e z_{i2}. Isso ocorre porque países com melhores tenistas tendem a ter um melhor desempenho na Copa Davis, resultando em um melhor *ranking* nacional.

Tabela 2.7 – Coeficientes de correlação de Pearson das variáveis transformadas dos dados de tênis da Copa Davis.

	w_i	z_{i1}	z_{i2}	z_{i3}	z_{i4}
w_i	1,00	0,49	0,42	0,39	0,06
z_{i1}		1,00	0,53	0,89	0,38
z_{i2}			1,00	0,50	0,91
z_{i3}				1,00	0,35
z_{i4}					1,00

Os critérios de seleção AIC e BIC avaliam o melhor ajuste do modelo de regressão proposto para os dados, considerando os menores valores. Desconsideramos as covariáveis z_{i3} e z_{i4} por apresentarem altíssima correlação com as covariáveis z_{i1} e z_{i2}, respectivamente e apresentarem menor correlação com a resposta que essas últimas. A Tabela 4.3 apresenta os valores de AIC e BIC, para os quatro possíveis modelos que envolvem z_{i1} e z_{i2}. Nota-se que o modelo M_2, apenas com a covariável z_{i1}, possui os menores valores, sendo, portanto, o modelo selecionado.

Tabela 2.8 – Critérios de seleção de modelos ajustados com dados de tênis da Copa Davis.

Modelo	AIC	BIC
$M_1 : g(p_i) = \beta_0$	$-8{,}5597$	$-8{,}0748$
$M_2 : g(p_i) = \beta_0 + \beta_1 z_{i1}$	$-11{,}8814$	$-10{,}9116$
$M_3 : g(p_i) = \beta_0 + \beta_1 z_{i2}$	$-10{,}0412$	$-9{,}0714$
$M_4 : g(p_i) = \beta_0 + \beta_1 z_{i1} + \beta_1 z_{i2}$	$-10{,}5070$	$-9{,}0523$

A Tabela 2.9 apresenta as estimativas dos parâmetros, os erros padrões, os valores observados da estatística de teste e o valores de p para testes sobre os parâmetros do modelo considerado. Observa-se que a diferença entre as médias dos *rankings* (z_{i1}) tem um efeito significativo na probabilidade do país A vencer um confronto com o país B. Além disso, podemos usar a razão de chances para interpretar as estimativas dos parâmetros do modelo ajustado. Estima-se, por exemplo, que a chance do país com melhor *ranking* vencer cada jogo aumenta em 0,67% $((\exp(0{,}0067) - 1) \times 100\%)$ para cada aumento de uma unidade na variável que representa a diferença entre as médias dos *rankings*.

Tabela 2.9 – Estimativas de máxima verossimilhança do modelo ajustado com dados de tênis da Copa Davis.

Parâmetro	Estimativa	Erro Padrão	Estatística Z	Valor-p
β_0	$-0{,}3951$	0,3009	$-0{,}7203$	0,4714
β_1	0,0067	0,0036	2,0153	0,0439

Para análise de diagnóstico utilizamos o gráfico de probabilidade normal com envelope simulado para o resíduo quantílico aleatorizado obtido a partir de w_i, apresentado na Figura 2.1. Todos os pontos estão dentro do envelope, não havendo portanto indícios de falta de ajuste do modelo de regressão proposto.

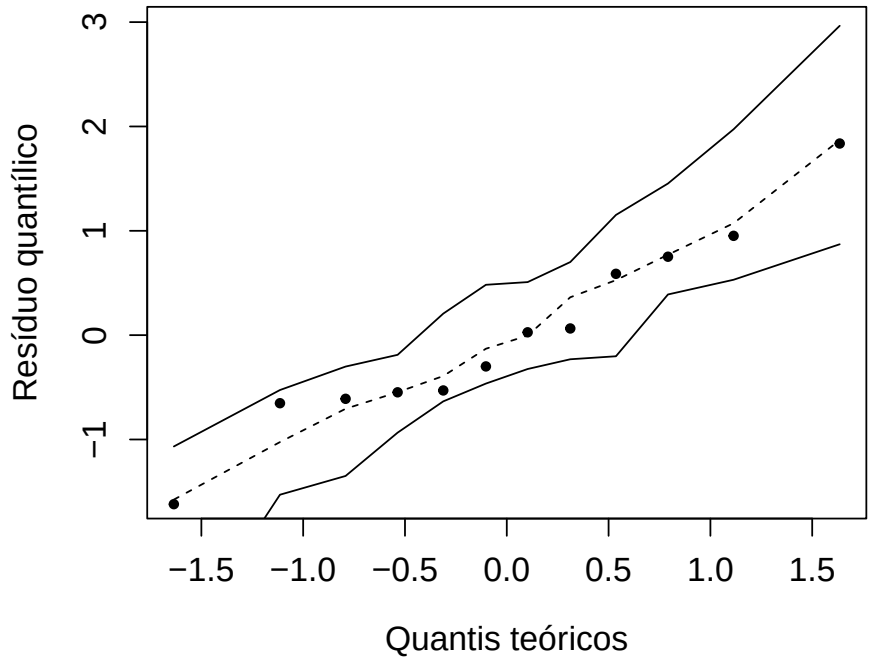


Figura 2.1 – Gráfico normal de probabilidade com envelope simulado para o modelo ajustado com os dados da Copa Davis

Apesar de o envelope indicar um ajuste satisfatório do modelo, este pode não ser adequado. Isso ocorre porque o modelo considerado neste capítulo assume que a probabilidade de um país vencer uma determinada partida é a mesma para todas as partidas do confronto. Na prática, entretanto, diferentes jogadores podem disputar as partidas de um confronto, de modo que essa probabilidade não se mantém constante ao longo do confronto.

2.6.2 Dados de tênis de mesa

O tênis de mesa, jogo popular entre os esportes divididos em *sets*, também chamado informalmente de *ping-pong* teve origem na Inglaterra como uma adaptação do tênis de quadra. Em 1921, foi criada uma Associação de Tênis de Mesa e, logo em seguida, em 1926 foi oficializada a Federação Internacional de Tênis de Mesa (IFTT). Os campeonatos mundiais, são disputados nas categorias individuais masculina e feminina, duplas masculina, feminina e mistas. Usualmente, em torneios nacionais as partidas de tênis de mesa são disputadas em melhor de

cinco *sets*, ganha a partida o jogador ou dupla que vence três *sets*. Em torneios internacionais, as partidas são disputadas em melhor de sete *sets*, ganha a partida o jogador ou equipe que vence quatro *sets*. Para ganhar um *set*, é precisa somar 11 pontos, ou em caso de empate vence o primeiro a abrir uma vantagem de dois pontos sobre o seu adversário.

O conjunto de dados foi obtido dos resultados das partidas do Campeonato Mundial de Tênis de Mesa de 2023, realizado em Durban na África do Sul nos dias 21 a 28 do mês de maio. O banco de dado foi extraído da base de dados do *Tênis de Mesa Mundial* da IFTT, disponível no site: <https://worldtabletennis.com/home>. Os dados escolhidos foram da competição individual feminino, disputada na versão melhor de sete *sets*, na qual vence a partida a jogadora que ganhar quatro *sets*. Selecionamos apenas a primeira rodada do torneio, totalizando 62 partidas, desta forma evitamos a dependência entre as observações, não usando os resultados das rodadas seguintes do campeonato. Como uma partida foi ganha de W.O., ou seja, teve uma ausência (ou impedimento) na disputa, totalizamos $n = 61$ observações no conjunto de dados. Este conjunto de dados contém os placares das partidas, nome, idade e *ranking* mundial atual da jogadora, respectivamente. Porém, para o uso da covariável *ranking* mundial, iremos obter os dados da IFTT disponível de forma gratuita no site oficial: <https://www.ittf.com/>, considerando os *rankings* mundiais das jogadoras na data anterior ao campeonato mundial, ou seja, na semana 20 disponível dia 16 de maio de 2023.

A Tabela 2.10 apresenta a organização dos dados, mostrando as observações iniciais e finais da base de dados.

Tabela 2.10 – Observações iniciais e finais do Campeonato Mundial de Tênis de Mesa 2023.

id	Age_A	Ranking_A	Placar_A	Placar_B	Ranking_B	Age_B
1	44	30	4	1	180	21
2	28	33	4	3	101	23
3	29	28	4	2	144	18
4	17	65	4	3	210	37
5	15	36	4	1	152	33
⋮	⋮	⋮	⋮	⋮	⋮	⋮
58	26	122	4	2	162	27
59	27	121	4	0	211	21
60	29	12	3	4	87	30
61	26	20	4	0	184	24
62	29	46	0	4	49	27

As variáveis de interesse são : $(x, y)-$ “placar A e placar B”, considerados como variáveis resposta, enquanto as variáveis “Ranking_A” e “Ranking_B” são consideradas como preditoras. O banco de dados foi reorganizado de modo que a jogadora com o melhor *ranking* foi selecionada ao longo do eixo x, na base de dados. A Tabela 2.11 apresenta a distribuição de frequência da variável resposta bivariada. Observamos que os placares “quatro a zero” e “quatro a um”, foram os mais frequentes nos dados, mostrando o melhor desempenho das jogadoras A, conforme

esperado já que elas possuem melhor *ranking*.

Tabela 2.11 – Distribuição de frequências para o resultado dos placares dos dados de tênis de mesa.

(x,y)	Freq. absoluta	Freq. relativa
(4,0)	23	0,37
(4,1)	13	0,21
(4,2)	10	0,16
(4,3)	4	0,07
(0,4)	4	0,07
(1,4)	1	0,02
(2,4)	5	0,08
(3,4)	1	0,02
Total	61	1

Para avaliar a seleção da covariável que apresente correlação com a variável resposta, faremos uso de transformações nas variáveis. Definimos as seguintes variáveis:

- $w_i = x_i - y_i$ a diferença dos placares;
- z_{i1} = ranking_B - ranking_A;
- z_{i2} = ranking_B/ranking_A;
- $z_{i3} = \log(z_{i1})$;
- $z_{i4} = \log(z_{i2})$.

A Tabela 2.12 apresenta os coeficientes de correlação de Pearson das variáveis transformadas. Observamos que a diferença entre os *rankings* (variável z_{i1}) possui uma correlação mais forte com a diferença da variável resposta. Além disso, as variáveis z_{i3} e z_{i4} também apresentam correlações semelhantes. No entanto, optamos por considerar a variável explicativa z_{i1} na análise, uma vez que sua interpretação é mais direta, já que as demais utilizam transformações logarítmicas.

Tabela 2.12 – Coeficientes de correlação de Pearson das variáveis transformadas dos dados de tênis de mesa.

Variáveis	w_i	z_{i1}	z_{i2}	z_{i3}	z_{i4}
w_i	1,00	0,44	0,35	0,43	0,43
z_{i1}		1,00	0,46	0,84	0,66
z_{i2}			1,00	0,38	0,88
z_{i3}				1,00	0,62
z_{i4}					1,00

Para este conjunto de dados assumimos o modelo de regressão proposto em (2.10) para modelar a probabilidade p_i da i-ésima jogadora A, com $i = 1, \ldots, n$, tais que

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_{i1}.$$

A Tabela 2.13 apresenta as estimativas de máxima verossimilhança dos parâmetros, os erros padrões, os valores observados da estatística de teste e seus respectivos valores de p sobre os parâmetros do modelo ajustado.

Tabela 2.13 – Estimativas de máxima verossimilhança do modelo ajustado com os dados de tênis de mesa.

Parâmetro	Estimativa	Erro Padrão	Estatística Z	Valor-p
β_0	$-0{,}0910$	0,1960	$-0{,}4640$	0,6426
β_1	0,0116	0,0023	5,1006	$<0{,}0001$

Com o objetivo de verificar possíveis desvios das suposições feitas sobre o modelo, construímos um gráfico de probabilidade normal com envelope simulado para os resíduos quantílicos aleatorizados, obtidos a partir de w_i. Pode-se observar, pela Figura 2.2, um ajuste insatisfatório, com muitos pontos próximos de zero fora dos limites do envelope. Além disso, os 7 menores resíduos também estão fora do envelope. A proporção de pontos fora dos limites do envelope é de 37,7%. Isso indica um ajuste inadequado aos dados, sugerindo a fuga da suposição de independência entre os *sets*.

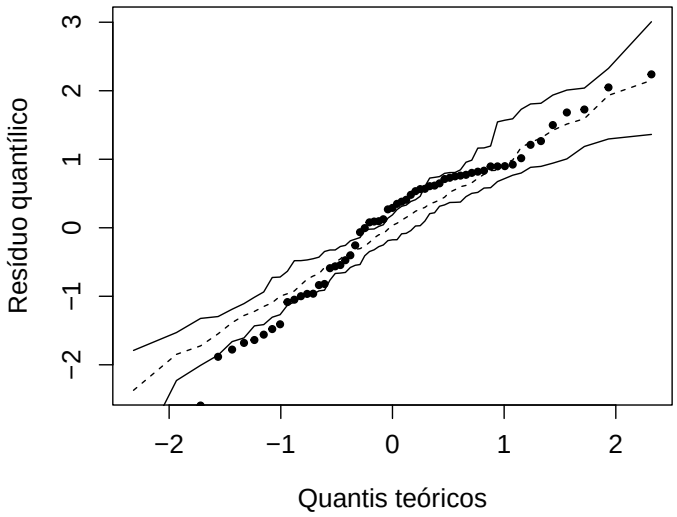


Figura 2.2 – Gráfico normal de probabilidade com envelope simulado do modelo ajustado com os dados de tênis de mesa

O fato de o envelope indicar que o modelo é inadequado aos dados não implica necessariamente que a suposição de independência entre os *sets* não seja atendida. Contudo, para este

conjunto de dados, tal suposição pode não ser razoável, uma vez que é natural supor que, se uma jogadora vence um *set*, a probabilidade de ela vencer o próximo aumenta. Assim, no Capítulo 3, será proposta uma distribuição de probabilidade que considere a dependência entre os *sets*, e, no Capítulo 4, será apresentado um modelo de regressão que assume que a variável resposta segue a distribuição de probabilidade introduzida no Capítulo 3.

CAPÍTULO

3

DISTRIBUIÇÃO DE PROBABILIDADE CONSIDERANDO DEPENDÊNCIA ENTRE OS *SETS*

No tênis, no tênis de mesa e no vôlei, é razoável imaginar que a probabilidade de um jogador ou time vencer um *set* é maior quando ele venceu o *set* anterior do que quando não venceu. Esse fenômeno pode ser explicado pela continuidade de um bom desempenho e pelo aumento de confiança adquirida ao vencer um *set*, fatores que influenciam positivamente o desempenho no *set* subsequente. Além disso, a vitória em um *set* pode proporcionar uma vantagem psicológica sobre o adversário, aumentando ainda mais as chances de vitória no próximo *set*. Esse efeito pode ser observado empiricamente, por exemplo, no tênis, onde o jogador de melhor *ranking* vence uma maior proporção de vezes o segundo *set* quando vence o primeiro, comparado a quando não vence.

Consideramos assim nesse capítulo, o caso em que a probabilidade de um jogador vencer um *set* **não é independente** do resultado no *set* anterior. Para mensurar essa dependência entre os *sets*, vamos propor uma cadeia de Markov em dois estados. Seja $\{W_1, W_2, ..., \}$ uma cadeia de Markov homogênea no tempo com valores $\{0, 1\}$, definidos por

$P(W_0 = 0) = p_0 \equiv$ probabilidade do jogador A vencer o primeiro *set*;
$P(W_0 = 1) = p_1 = 1 - p_0 \equiv$ probabilidade do jogador B vencer o primeiro *set*;
$P(W_{t+1} = 0|W_t = 0) = p_{00} \equiv$ probabilidade de A vencer o *set* dado que ganhou o anterior;
$P(W_{t+1} = 1|W_t = 0) = 1 - p_{00} \equiv$ probabilidade de B vencer o *set* dado que A ganhou o anterior;
$P(W_{t+1} = 1|W_t = 1) = p_{11} \equiv$ probabilidade de B vencer o *set* dado que ganhou o anterior;
$P(W_{t+1} = 0|W_t = 1) = 1 - p_{11} \equiv$ probabilidade de A vencer o *set* dado que B ganhou o anterior,

em que $i = 0, 1, 2, \ldots$.

Na prática, é razoável supor que $p_0 > 0,5$, especialmente porque o jogador A possui um *ranking* superior. Além disso, espera-se que $p_{00} > p_0$ e, de forma equivalente, $p_{11} > p_1$, já que é razoável imaginar que a probabilidade de um jogador vencer um *set* aumente quando ele venceu o *set* anterior. Finalmente, como última relação esperada, $p_{00} > p_{11}$, uma vez que o jogador A tem um *ranking* melhor.

3.1 Distribuição de probabilidade bivariada proposta

Para a construção da função de probabilidade, considerando dependência Markoviana entre os *sets*, iremos apresentar os seguintes exemplos.

Considere um jogo na versão melhor de três *sets*, na qual ganha uma partida o jogador que vencer dois *sets*. Por meio da cadeia de Markov em dois estados homogêneos, podemos calcular, por exemplo, a probabilidade do jogador B vencer a partida em três *sets*. Na Tabela 3.1, apresentamos as possíveis formas do jogador B vencer a partida em três *sets*.

Tabela 3.1 – Evento: jogador B vencer a partida em três *sets*.

Eventos	1º *set*	2º *set*	3º *set*	probabilidade
1	A	B	B	$p_0(1-p_{00})p_{11}$
2	B	A	B	$p_1(1-p_{11})(1-p_{00})$

Dessa forma, a probabilidade do jogador B vencer a partida em três *sets* é dada por

$$P(X=1,Y=2) = p_0(1-p_{00})p_{11} + p_1(1-p_{11})(1-p_{00}).$$

Agora, considere um jogo na versão melhor de cinco *sets*, em que o vencedor da partida é o jogador que ganha três *sets*. Podemos, então, calcular a probabilidade de o jogador A vencer a partida em cinco *sets*. Na Tabela 3.2, apresentamos as possíveis formas de o jogador A vencer a partida em cinco *sets*.

Tabela 3.2 – Evento: jogador A vencer a partida em cinco *sets*.

Eventos	1º*set*	2º*set*	3º*set*	4º*set*	5º*set*	probabilidade
1	A	A	B	B	A	$p_0p_{00}p_{11}(1-p_{00})(1-p_{11})$
2	A	B	B	A	A	$p_0p_{00}p_{11}(1-p_{00})(1-p_{11})$
3	A	B	A	B	A	$p_0(1-p_{00})^2(1-p_{11})^2$
4	B	B	A	A	A	$p_1p_{00}^2p_{11}(1-p_{11})$
5	B	A	B	A	A	$p_1p_{00}(1-p_{11})^2(1-p_{00})$
6	B	A	A	B	A	$p_1p_{00}(1-p_{11})^2(1-p_{00})$

Observe que, para alguns eventos as probabilidades são iguais. Portanto, a probabilidade

de o jogador A vencer a partida em cinco *sets* é dada por

$$\begin{aligned} P(X=3,Y=2) &= 2p_0p_{00}p_{11}(1-p_{00})(1-p_{11})+2p_1p_{00}(1-p_{11})^2(1-p_{00})+ \\ &\quad p_0(1-p_{00})^2(1-p_{11})^2+p_1p_{00}^2p_{11}(1-p_{11}). \end{aligned}$$

De forma análoga, podemos calcular as probabilidades para um jogo na versão melhor de sete *sets*, em que o jogador que ganhar quatro *sets* vence a partida. Podemos exemplificar o cálculo considerando o seguinte evento de interesse: o jogador A vencer a partida em seis *sets*. Na Tabela 3.3, apresentamos as possibilidades de o jogador A vencer a partida em seis *sets*.

Tabela 3.3 – Evento: jogador A vencer a partida em seis *sets*.

Eventos	$1^o set$	$2^o set$	$3^o set$	$4^o set$	$5^o set$	$6^o set$	probabilidade
1	A	B	B	A	A	A	$p_0(1-p_{00})p_{11}(1-p_{11})p_{00}^2$
2	A	A	B	B	A	A	$p_0p_{00}(1-p_{00})p_{11}(1-p_{11})p_{00}$
3	A	A	A	B	B	A	$p_0p_{00}^2(1-p_{00})p_{11}(1-p_{11})$
4	A	B	A	A	B	A	$p_0(1-p_{00})^2(1-p_{11})^2p_{00}$
5	A	A	B	A	B	A	$p_0p_{00}(1-p_{00})^2(1-p_{11})^2$
6	A	B	A	B	A	A	$p_0(1-p_{00})^2(1-p_{11})^2p_{00}$
7	B	B	A	A	A	A	$p_1p_{11}(1-p_{11})p_{00}^3$
8	B	A	A	A	B	A	$p_1(1-p_{11})^2p_{00}^2(1-p_{00})$
9	B	A	A	B	A	A	$p_1(1-p_{11})^2p_{00}(1-p_{00})p_{00}$
10	B	A	B	A	A	A	$p_1(1-p_{11})^2(1-p_{00})p_{00}^2$

Observe novamente que alguns eventos possuem as mesmas probabilidades. Assim, a probabilidade de o jogador A vencer a partida em seis *sets* é determinada por

$$\begin{aligned} P(X=4,Y=2) &= p_0\{3p_{00}^2p_{11}(1-p_{00})(1-p_{11})+3p_{00}^2(1-p_{00})^2(1-p_{11})^2\}+ \\ &\quad p_1\{p_{00}^3p_{11}(1-p_{11})+3p_{00}^2(1-p_{00})(1-p_{11})^2\}. \end{aligned}$$

No tênis, os torneios de Grand Slam são as principais competições do ano. Há 4 torneios de Grand Slam no ano e eles são realizados na Austrália, França, Inglaterra e Estados Unidos. Como nesses torneios os jogos são disputados na versão melhor de cinco *sets*, nesse estudo será considerada a função de probabilidade para dados de jogos com essa característica.

Portanto, fixando $k=3$, temos que $(X,Y)^\top$ é um vetor aleatório bidimensional discreto, onde o máximo de X e Y é igual a k, que é o número de *sets* necessário para um jogador vencer a partida, logo, $x,y=0,\ldots,3$, e o $\min(X,Y)=0,\cdots,k-1$. Considerando que a dependência entre os *sets* é mensurada pela cadeia de Markov homogênea em dois estados, a função de probabilidade é dada por,

$$P(X=i,Y=j)=a_{ij}, \tag{3.1}$$

em que

$$\begin{aligned}
a_{30} &= p_0 p_{00}^2, \\
a_{31} &= p_1 p_{00}^2(1-p_{11}) + 2p_0 p_{00}(1-p_{00})(1-p_{11}), \\
a_{32} &= 2p_0 p_{00} p_{11}(1-p_{00})(1-p_{11}) + 2p_1 p_{00}(1-p_{11})^2(1-p_{00}) + \\
&\quad p_0(1-p_{00})^2(1-p_{11})^2 + p_1 p_{00}^2 p_{11}(1-p_{11}), \\
a_{03} &= p_1 p_{11}^2, \\
a_{13} &= p_0 p_{11}^2(1-p_{00}) + 2p_1 p_{11}(1-p_{00})(1-p_{11}), \\
a_{23} &= 2p_0 p_{11}(1-p_{11})(1-p_{00})^2 + \\
&\quad 2p_1 p_{00} p_{11}(1-p_{00})(1-p_{11}) + \\
&\quad p_0 p_{00} p_{11}^2(1-p_{00}) + p_1(1-p_{11})^2(1-p_{00})^2,
\end{aligned}$$

em que $0 < p_0, p_{00}, p_{11} < 1$ são os parâmetros do modelo.

Para gerar números aleatórios dessa distribuição, utilizamos o seguinte procedimento numérico baseado na função de probabilidade proposta:

Algoritmo 1 – Algoritmo para gerar valores pseudo-aleatórios.

1: Definir valores para o vetor de parâmetros $\boldsymbol{\theta} = (p_0, p_{00}, p_{11})$;
2: Calcular as probabilidades $P(X = i, Y = j) = a_{ij}$, com $i, j = 0, \ldots, 3$;
3: Calcular o vetor de soma cumulativas das probabilidade a_{ij}, definido por q_l, $l = 1, \ldots, 6$;
4: Gerar um valor com distribuição uniforme, $u \sim (0, 1)$;
5: Se $u < q_1$ então $(x = 3, y = 0)$;
6: Se $q_1 < u < q_2$ então $(x = 3, y = 1)$;
7: Se $q_2 < u < q_3$ então $(x = 3, y = 2)$;
8: Se $q_3 < u < q_4$ então $(x = 0, y = 3)$;
9: Se $q_4 < u < q_5$ então $(x = 1, y = 3)$;
10: Se $q_5 < u < q_6$ então $(x = 2, y = 3)$;
11: Repetir o processo n vezes.

3.1.1 Estimação dos parâmetros

Seja $(x_1, y_1), \ldots, (x_n, y_n)$ os valores observados de uma amostra das variáveis aleatórias bivariadas $(X_1, Y_1), \ldots, (X_n, Y_n)$ de um vetor que possui função de probabilidade dada por $P(X = x, Y = y|\boldsymbol{\theta})$ em (3.1), com $\boldsymbol{\theta} = (p_0, p_{00}, p_{11})$. A função de verossimilhança para $\boldsymbol{\theta}$ é dada por

$$\begin{aligned}
L(\boldsymbol{\theta}; (\mathbf{x}, \mathbf{y})) &= \prod_{i=1}^{n} P(X = x_i, Y = y_i) \\
&= \prod_{i=1}^{n} \Big[a_{30}^{I_{\{3,0\}}(x_i,y_i)} \times a_{31}^{I_{\{3,1\}}(x_i,y_i)} \times \\
&\quad a_{32}^{I_{\{3,2\}}(x_i,y_i)} \times a_{03}^{I_{\{0,3\}}(x_i,y_i)} \times \\
&\quad a_{13}^{I_{\{1,3\}}(x_i,y_i)} \times a_{23}^{I_{\{2,3\}}(x_i,y_i)} \Big].
\end{aligned}$$

Então, o logaritmo da função de verossimilhança é dada por

$$
\begin{aligned}
\ell(\boldsymbol{\theta};(\mathbf{x},\mathbf{y})) &= \sum_{i=1}^{n} \log(L(\boldsymbol{\theta};(\mathbf{x},\mathbf{y}))) \\
&= \sum_{i=1}^{n} I_{\{3,0\}}(x_i,y_i)\log\left[p_0 p_{00}^2\right] + \\
&\quad \sum_{i=1}^{n} I_{\{3,1\}}(x_i,y_i)\log\left[p_1 p_{00}^2(1-p_{11}) + p_0(2p_{00}(1-p_{00})(1-p_{11}))\right] + \\
&\quad \sum_{i=1}^{n} I_{\{3,2\}}(x_i,y_i)\log\left[2\times p_0 p_{00} p_{11}(1-p_{00})(1-p_{11}) + \right. \\
&\quad \left. 2\times p_1 p_{00}(1-p_{11})^2(1-p_{00}) + p_0(1-p_{00})^2(1-p_{11})^2 + p_1 p_{00}^2 p_{11}(1-p_{11})\right] + \\
&\quad \sum_{i=1}^{n} I_{\{0,3\}}(x_i,y_i)\log\left[p_1 p_{11}^2\right] + \qquad (3.2) \\
&\quad \sum_{i=1}^{n} I_{\{1,3\}}(x_i,y_i)\log\left[p_0 p_{11}^2(1-p_{00}) + 2\times p_1 p_{11}(1-p_{00})(1-p_{11})\right] + \\
&\quad \sum_{i=1}^{n} I_{\{2,3\}}(x_i,y_i)\log\left[2\times p_0 p_{11}(1-p_{11})(1-p_{00})^2 + \right. \\
&\quad \left. 2\times p_1 p_{00} p_{11}(1-p_{00})(1-p_{11}) + p_0 p_{00} p_{11}^2(1-p_{00}) + p_1(1-p_{11})^2(1-p_{00})^2\right],
\end{aligned}
$$

em que I é a função indicadora, definida por

$$
I_{\{i,j\}}(x,y) = \begin{cases} 1, & \text{se} \quad x=i, y=j; \\ 0, & \text{caso contrário.} \end{cases}
$$

Para obter os estimadores, utilizamos o método da máxima verossimilhança. Contudo, nesse cenário, a derivada do logaritmo da função de verossimilhança em (3.3) não possui uma solução analítica explícita. Por isso, empregamos o procedimento de aproximação numérica quase-Newton BFGS, utilizando a função *optim* na linguagem de programação *R*.

3.1.2 Estudos de simulação

Para exemplificar a metodologia apresentada, simulamos dados associados a jogos no formato melhor de cinco *sets* e ajustamos esses dados utilizando a função de probabilidade proposta, considerando diferentes valores dos parâmetros e usando o algoritmo para geração de valores pseudo-aleatórios da variável resposta (Algoritmo 1). O estudo de simulação de Monte Carlo utilizou 5000 réplicas. A função de probabilidade proposta possui três parâmetros, $\boldsymbol{\theta} = (p_0, p_{00}, p_{11})$ e consideramos quatro diferentes valores para o vetor $\boldsymbol{\theta}$. Os tamanhos amostrais considerados foram $n = 20, 40, 60, 100, 500, 1000$.

A escolha dos valores iniciais foi realizada com base nas definições e interpretações de cada parâmetro. Por exemplo, p_0 representa a probabilidade de o jogador A vencer o primeiro

set; portanto, é razoável propor como valor inicial a razão entre as proporções amostrais que favorecem o jogador A. As duas estimativas iniciais seguintes foram definidas de modo a preservar a coerência com a interpretação prática dos parâmetros, conforme já descrito anteriormente, assumindo-se $p_{00} > p_0$ e $p_{11} > p_1$. Assim, a segunda estimativa inicial foi obtida, por exemplo, a partir da razão de proporções que considera a média amostral de x apenas nos jogos em que o jogador A foi vencedor, de modo a refletir sua vantagem; de forma análoga, foi definida a estimativa inicial para o terceiro parâmetro. Dessa maneira, as estimativas iniciais utilizadas no início do procedimento numérico foram: $\widehat{\boldsymbol{\theta}}_0 = \left(\frac{\overline{x}}{\overline{x}+\overline{y}}, \frac{\overline{x}_{I_3(x_i)}}{\overline{x}_{I_3(x_i)}+\overline{y}}, \frac{\overline{y}_{I_3(y_i)}}{\overline{y}_{I_3(y_i)}+\overline{x}}\right)$, em que $\overline{x}_{I_3(x_i)}$ denota a média que considera somente os placares vencidos pelo jogador A, enquanto $\overline{y}_{I_3(y_i)}$ corresponde à média dos placares vencidos pelo jogador B. Para a análise do desempenho do estimador de máxima verossimilhança, foram calculadas, para cada tamanho amostral, as seguintes medidas: estimativa do viés e estimativa da raiz quadrada do erro quadrático médio, conforme definidas na Seção 2.5.

No estudo de simulação, os cenários foram escolhidos para caracterizar as seguintes probabilidades de vitória. No primeiro cenário, temos os seguintes valores dos parâmetros: $p_0 = 0,6$, $p_{00} = 0,7$, $p_{11} = 0,5$, o que indica que o jogador A tem uma probabilidade razoável de vencer o primeiro *set* e, uma vez que ganhou o *set* anterior, a probabilidade de vencer o próximo é relativamente alta. Por outro lado, o jogador B tem uma probabilidade moderada de vencer um *set* dado que ganhou o anterior. No segundo cenário, os valores dos parâmetros foram: $p_0 = 0,8$, $p_{00} = 0,9$, $p_{11} = 0,7$, indicando que o jogador A tem uma alta probabilidade de vencer o primeiro *set* e, uma vez que ganhou o *set* anterior, a probabilidade de vencer o próximo é muito alta. O jogador B ainda tem uma chance significativa de vencer um *set* dado que ganhou o anterior. No terceiro cenário, os parâmetros escolhidos foram: $p_0 = 0,6$, $p_{00} = 0,9$, $p_{11} = 0,8$, especificando que o jogador A tem uma probabilidade moderada de vencer o primeiro *set* e, uma vez que ganhou o *set* anterior, a probabilidade de vencer o próximo é muito alta. O jogador B tem uma boa chance de vencer um *set* dado que ganhou o anterior. No último cenário, os parâmetros foram determinados para corresponder a um cenário de independência de vitórias entre os *sets*, com os valores: $p_0 = 0,6$, $p_{00} = 0,6$, $p_{11} = 0,4$, indicando que o jogador A tem uma probabilidade moderada de vencer o primeiro *set*, e uma vez que ganhou o *set* anterior, a probabilidade de vencer o próximo também é moderada e a mesma. O jogador B tem uma chance um pouco menor de vencer o primeiro *set* e, uma vez que ganhou o *set* anterior, a probabilidade de vencer o próximo é a mesma que vencer o primeiro *set*.

Na Tabela 3.4, são apresentados os resultados obtidos a partir da simulação. Observamos que os estimadores de máxima verossimilhança dos parâmetros apresentam valores pequenos de viés, mesmo para tamanhos amostrais pequenos, e parecem ser não viciados para amostras grandes. Ao avaliar os diferentes cenários, constatamos que o primeiro e o quarto cenários apresentaram maiores valores de viés e REQM, em comparação com o segundo e o terceiro cenário. Além disso, de modo geral, o viés e o REQM dos estimadores diminuem à medida que o tamanho amostral aumenta, conforme esperado, evidenciando a propriedade de consistência

do estimador de máxima verossimilhança.

Tabela 3.4 – Estimativa dos viés e REQM (em parênteses) dos estimadores dos parâmetros do modelo proposto para os dados simulados com alguns valores de n e $\boldsymbol{\theta} = (p_0, p_{00}, p_{11})$.

n	p_0 $(0,60)$	p_{00} $(0,70)$	p_{11} $(0,50)$	p_0 $(0,80)$	p_{00} $(0,90)$	p_{11} $(0,70)$
20	0,013 (0,228)	0,004 (0,121)	0,014 (0,132)	0,003 (0,121)	0,000 (0,066)	−0,004 (0,154)
40	0,008 (0,168)	0,003 (0,088)	0,009 (0,095)	0,000 (0,086)	0,001 (0,046)	−0,001 (0,103)
60	0,005 (0,138)	0,002 (0,073)	0,007 (0,079)	0,000 (0,069)	0,000 (0,037)	−0,001 (0,083)
100	0,003 (0,109)	0,002 (0,057)	0,005 (0,061)	−0,001 (0,053)	0,000 (0,029)	−0,001 (0,065)
500	0,000 (0,052)	0,001 (0,026)	0,001 (0,028)	0,000 (0,024)	0,000 (0,013)	−0,001 (0,029)
1000	0,000 (0,037)	0,000 (0,019)	0,001 (0,020)	0,000 (0,017)	0,000 (0,009)	0,000 (0,020)
n	p_0 $(0,60)$	p_{00} $(0,90)$	p_{11} $(0,80)$	p_0 $(0,60)$	p_{00} $(0,60)$	p_{11} $(0,40)$
20	0,001 (0,128)	−0,001 (0,068)	−0,005 (0,107)	0,022 (0,289)	0,012 (0,139)	0,031 (0,139)
40	0,000 (0,092)	0,000 (0,048)	−0,002 (0,073)	0,011 (0,222)	0,009 (0,104)	0,020 (0,102)
60	0,000 (0,074)	0,000 (0,039)	−0,001 (0,059)	0,007 (0,187)	0,007 (0,087)	0,015 (0,086)
100	−0,001 (0,057)	0,000 (0,030)	−0,001 (0,046)	0,007 (0,153)	0,005 (0,070)	0,011 (0,068)
500	0,000 (0,026)	0,000 (0,013)	0,000 (0,020)	0,001 (0,076)	0,002 (0,033)	0,003 (0,032)
1000	0,000 (0,018)	0,000 (0,009)	0,000 (0,014)	0,001 (0,054)	0,001 (0,024)	0,001 (0,023)

Nesta seção não apresentamos uma aplicação, uma vez que, na prática, em jogos divididos em *sets*, o vetor de probabilidades associado à variável resposta varia em função de variáveis preditoras. Por esse motivo, no próximo capítulo será realizada uma aplicação que considera essas características.

CAPÍTULO

4

MODELOS DE REGRESSÃO CONSIDERANDO DEPENDÊNCIA ENTRE OS *SETS*

Na Seção 3.1 do Capítulo 3, definimos a variável resposta (X,Y) como um vetor aleatório bivariado discreto que representa o resultado de partidas de jogos divididos em *sets*, considerando a dependência entre os *sets* por meio de cadeias de Markov, cujos parâmetros são $0 < p_0, p_{00}, p_{11} < 1$.

No contexto de regressão, sejam $(X_1,Y_1),\ldots,(X_n,Y_n)$ variáveis aleatórias independentes, em que cada (X_i,Y_i) segue a distribuição definida em (3.1). Suponha que $(x_1,y_1),\ldots,(x_n,y_n)$ sejam observações de $(X_1,Y_1),\ldots,(X_n,Y_n)$, respectivamente. Os parâmetros da variável resposta são, portanto, definidos como $\boldsymbol{p}_0 = (p_{01},\ldots,p_{0n})^\top$, $\boldsymbol{p}_{00} = (p_{001},\ldots,p_{00n})^\top$ e $\boldsymbol{p}_{11} = (p_{111},\ldots,p_{11n})^\top$, os quais satisfazem as relações:

$$\begin{cases} g_1(p_{0i}) = \boldsymbol{z}_{i1}^\top \boldsymbol{\beta}_1 = \eta_{i1} \\ g_2(p_{00i}) = \boldsymbol{z}_{i2}^\top \boldsymbol{\beta}_2 = \eta_{i2} \\ g_3(p_{11i}) = \boldsymbol{z}_{i3}^\top \boldsymbol{\beta}_3 = \eta_{i3}, \end{cases} \tag{4.1}$$

em que $\boldsymbol{\beta}_1 = (\beta_{01},\beta_{11},\ldots,\beta_{J_11})^\top$, $\boldsymbol{\beta}_2 = (\beta_{02},\beta_{12},\ldots,\beta_{J_22})^\top$ e $\boldsymbol{\beta}_3 = (\beta_{03},\beta_{13},\ldots,\beta_{J_33})^\top$ são os vetores de parâmetros desconhecidos. Os vetores de constantes que representam os valores das variáveis explicativas são dados por $\boldsymbol{z}_{i1} = (1,z_{i21}\ldots,z_{iJ_11})^\top$, $\boldsymbol{z}_{i2} = (1,z_{i22}\ldots,z_{iJ_22})^\top$ e $\boldsymbol{z}_{i3} = (1,z_{i23}\ldots,z_{iJ_33})^\top$. As funções de ligação $g_l(.)(l = 1,2,3)$, são estritamente monótonas e duas vezes diferenciáveis, relacionando os componentes sistemáticos aos componentes aleatórios, as principais funções utilizadas são apresentadas na Seção 2.2.

4.1 Estimação dos parâmetros

As estimativas dos parâmetros do modelo são obtidas pelo método da máxima verossimilhança, que consiste em maximizar o logaritmo da função de verossimilhança conjunta dos dados. O logaritmo da função de verossimilhança do modelo é dada por:

$$\begin{aligned}\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})) &= \sum_{i=1}^{n} I_{\{3,0\}}(x_i,y_i)\log(a_{30i}) + \sum_{i=1}^{n} I_{\{3,1\}}(x_i,y_i)\log(a_{31i}) + \\ &\sum_{i=1}^{n} I_{\{3,2\}}(x_i,y_i)\log(a_{32i}) + \sum_{i=1}^{n} I_{\{0,3\}}(x_i,y_i)\log(a_{03i}) + \\ &\sum_{i=1}^{n} I_{\{1,3\}}(x_i,y_i)\log(a_{13i}) + \sum_{i=1}^{n} I_{\{2,3\}}(x_i,y_i)\log(a_{23i})\,. \end{aligned} \tag{4.2}$$

Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)^\top$ o vetor com dimensão total $s := J_1 + J_2 + J_3 + 3$ de parâmetros do modelo (4.1). O vetor escore é obtido derivando-se o logaritmo da função de verossimilhança (4.2) em relação a cada parâmetro. Após manipulações algébricas (ver Apêndice A.2), obtém-se:

$$U(\boldsymbol{\theta}) = \left(U_{\boldsymbol{\beta}_1}^\top(\boldsymbol{\theta}), U_{\boldsymbol{\beta}_2}^\top(\boldsymbol{\theta}), U_{\boldsymbol{\beta}_3}^\top(\boldsymbol{\theta})\right)^\top. \tag{4.3}$$

As componentes do vetor escore (4.3) são dadas por:

$$\begin{aligned} U_{\boldsymbol{\beta}_1}(\boldsymbol{\theta}) &= \boldsymbol{Z}_1^\top \boldsymbol{D}_{p_0} \boldsymbol{L}_{p_0}, \\ U_{\boldsymbol{\beta}_2}(\boldsymbol{\theta}) &= \boldsymbol{Z}_2^\top \boldsymbol{D}_{p_{00}} \boldsymbol{L}_{p_{00}}, \\ U_{\boldsymbol{\beta}_3}(\boldsymbol{\theta}) &= \boldsymbol{Z}_3^\top \boldsymbol{D}_{p_{11}} \boldsymbol{L}_{p_{11}}, \end{aligned}$$

em que, para $l = 1,2,3$, as matrizes $\boldsymbol{Z}_l$ são de dimensão $n \times J_l$, com i-ésima linha dada por $\boldsymbol{z}_{il}^\top$ e $\boldsymbol{D}_{p_0} = \operatorname{diag}\{d_{p_{01}},\ldots,d_{p_{0n}}\}$; $\boldsymbol{D}_{p_{00}} = \operatorname{diag}\{d_{p_{001}},\ldots,d_{p_{00n}}\}$; $\boldsymbol{D}_{p_{11}} = \operatorname{diag}\{d_{p_{111}},\ldots,d_{p_{11n}}\}$; $\boldsymbol{L}_{p_0} = (l_{p_{01}},\ldots,l_{p_{0n}})^\top$; $\boldsymbol{L}_{p_{00}} = (l_{p_{001}},\ldots,l_{p_{00n}})^\top$; $\boldsymbol{L}_{p_{11}} = (l_{p_{111}},\ldots,l_{p_{11n}})^\top$, em que os elementos das matrizes e vetores são definidos no Apêndice A.2.

A matriz de informação de Fisher é definida a partir da função escore (4.3) e pode ser escrita como:

$$\boldsymbol{J}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} & \boldsymbol{J}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_2} & \boldsymbol{J}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_3} \\ \boldsymbol{J}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_1} & \boldsymbol{J}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2} & \boldsymbol{J}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_3} \\ \boldsymbol{J}_{\boldsymbol{\beta}_3\boldsymbol{\beta}_1} & \boldsymbol{J}_{\boldsymbol{\beta}_3\boldsymbol{\beta}_2} & \boldsymbol{J}_{\boldsymbol{\beta}_3\boldsymbol{\beta}_3} \end{pmatrix}, \tag{4.4}$$

com componentes dadas por:

$$
\begin{aligned}
\boldsymbol{J}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} &= \boldsymbol{Z}_1^\top \boldsymbol{F}_{p_0} \boldsymbol{L}_{p_0} \boldsymbol{L}_{p_0}^\top \boldsymbol{Z}_1, \\
\boldsymbol{J}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2} &= \boldsymbol{Z}_2^\top \boldsymbol{F}_{p_{00}} \boldsymbol{L}_{p_{00}} \boldsymbol{L}_{p_{00}}^\top \boldsymbol{Z}_2, \\
\boldsymbol{J}_{\boldsymbol{\beta}_3\boldsymbol{\beta}_3} &= \boldsymbol{Z}_3^\top \boldsymbol{F}_{p_{11}} \boldsymbol{L}_{p_{11}} \boldsymbol{L}_{p_{11}}^\top \boldsymbol{Z}_3, \\
\boldsymbol{J}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_2} &= \boldsymbol{Z}_1^\top \boldsymbol{F}_{p_0 p_{00}} \boldsymbol{L}_{p_0} \boldsymbol{L}_{p_{00}}^\top \boldsymbol{Z}_2, \\
\boldsymbol{J}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_3} &= \boldsymbol{Z}_1^\top \boldsymbol{F}_{p_0 p_{11}} \boldsymbol{L}_{p_0} \boldsymbol{L}_{p_{11}}^\top \boldsymbol{Z}_3, \\
\boldsymbol{J}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_3} &= \boldsymbol{Z}_2^\top \boldsymbol{F}_{p_{00} p_{11}} \boldsymbol{L}_{p_{00}} \boldsymbol{L}_{p_{11}}^\top \boldsymbol{Z}_3.
\end{aligned}
$$

em que, $\boldsymbol{F}_{p_0} = \text{diag}(f_{p_{0_1}}, \ldots, f_{p_{0_n}})$, $\boldsymbol{F}_{p_{00}} = \text{diag}(f_{p_{00_1}}, \ldots, f_{p_{00_n}})$, $\boldsymbol{F}_{p_{11}} = \text{diag}(f_{p_{11_1}}, \ldots, f_{p_{11_n}})$ e, por fim, as matrizes cruzadas definidas como $\boldsymbol{F}_{p_0 p_{00}} = \text{diag}(f_{p_0 p_{00_1}}, \ldots, f_{p_0 p_{00_n}})$, $\boldsymbol{F}_{p_0 p_{11}} = \text{diag}(f_{p_0 p_{11_1}}, \ldots, f_{p_0 p_{11_n}})$, $\boldsymbol{F}_{p_{00} p_{11}} = \text{diag}(f_{p_{00} p_{11_1}}, \ldots, f_{p_{00} p_{11_n}})$, em que os componentes dessas matrizes são definidos no Apêndice A.2.

Dada a inexistência de solução analítica explícita para o sistema de equações da função escore, os estimadores de máxima verossimilhança dos parâmetros do modelo, denotados por $\widehat{\boldsymbol{\theta}}$, são obtidos por meio de métodos numéricos. Em todos os estudos de simulação de Monte Carlo e na aplicação apresentada no Capítulo 4, foi empregado o método quase-Newton BFGS para a otimização numérica, de forma semelhante ao usado no Capítulo 2.

4.2 Análise de diagnóstico

A análise diagnóstico é parte fundamental para verificar a adequação do modelo proposto. Duas técnicas importantes são utilizadas, os resíduos para analisar o ajuste do modelo, e uma medida de influência global para a identificação de observações influentes no modelo.

4.2.1 Análise de resíduos

Assim como no modelo que assume independência entre os *sets*, também propomos o uso do resíduo quantílico aleatorizado no modelo que considera dependência Markoviana. Uma das vantagens do resíduo quantílico aleatorizado é que sua formulação apresenta a mesma estrutura em diferentes modelos. Dessa forma, a expressão do resíduo quantílico para o modelo de regressão considerado neste capítulo é análoga à apresentada na Seção 2.4, diferenciando-se apenas pela forma da função de distribuição acumulada, que é distinta nos dois modelos. Além disso, no modelo que incorpora dependência Markoviana entre os *sets*, o resíduo quantílico aleatorizado também segue, sob o modelo correto, uma distribuição assintoticamente normal padrão, o que é bastante conveniente para a avaliação do ajuste do modelo.

4.2.2 Influência global

O afastamento de verossimilhança é uma métrica utilizada para avaliar a influência global de uma ou mais observações em um modelo de regressão, indicando o quanto essa observação altera as estimativas dos coeficientes do modelo, conforme discutido em Cook, Peña e Weisberg (1982). Valores elevados dessa medida podem indicar a presença de pontos influentes que podem levar a conclusões inferenciais imprecisas do modelo. Trabalhos recentes (CORTÉS; CASTRO; GALLARDO, 2023; FABIO *et al.*, 2023) utilizam essa medida que pode ser expressa como uma definição análoga para o afastamento de verossimilhança em modelos de regressão, dada por

$$\mathrm{LD}_i = 2\left[\ell(\hat{\boldsymbol{\theta}} \mid (\boldsymbol{x},\boldsymbol{y})) - \ell(\hat{\boldsymbol{\theta}}_{(i)} \mid (\boldsymbol{x},\boldsymbol{y}))\right], \tag{4.5}$$

em que $\hat{\boldsymbol{\theta}}_{(i)}$ denota o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ obtido após a exclusão da i-ésima observação, e $\ell(\cdot)$ representa o logaritmo da função de verossimilhança definida em (4.2). O afastamento de verossimilhança quantifica a diferença entre as estimativas $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(i)}$ por meio da distância entre o logaritmo das respectivas funções de verossimilhança, permitindo avaliar o impacto das observações sobre os parâmetros estimados.

4.3 Estudos de simulação

Nesta seção, apresenta-se um estudo de simulação com o objetivo de avaliar a performance do estimador de máxima verossimilhança no modelo de regressão proposto em (4.1). As estimativas de máxima verossimilhança para os parâmetros do modelo foram obtidas e foram avaliadas as estimativas do viés e da raiz do erro quadrático médio, conforme definidos na Seção 2.5.

O conjunto de dados com distribuição bivariada proposta (3.1) foram simulados considerando $k = 3$, ou seja, simula um jogo na versão melhor de cinco *sets*, que ganha a partida um jogador (ou equipe) que vencer três *sets*. Foram avaliados quatro diferentes cenários, obtidos de quatro diferentes valores para os vetores de parâmetros. Duas covariáveis foram consideradas, com distribuições $Z_{i1} \sim U(0,1)$ e $Z_{i2} \sim U(0,1)$, $i = 1,\ldots,n$, e estas foram mantidas constantes em todas as S réplicas de Monte Carlo.

O estudo foi realizado por meio de simulação de Monte Carlo com $S = 5000$ réplicas, considerando os seguintes tamanhos amostrais: $n = 50, 100, 200, 500, 1000$. A função de ligação logito foi escolhida para o modelo de regressão, conforme citado na Seção 2.5. No primeiro cenário, os coeficientes da regressão foram $\beta_{01} = 1,2, \beta_{11} = 0,5, \beta_{21} = -0,5, \beta_{02} = 1,5, \beta_{12} = 0,5, \beta_{22} = -0,5, \beta_{03} = 0,2, \beta_{13} = -0,5, \beta_{23} = -0,5$, indicando que os melhores jogadores têm, em média, maiores probabilidades de ganhar o primeiro *set*, e possuem alta probabilidade de vitória em um *set* dado que venceu o anterior, e que as variáveis regressoras possuem um efeito fraco na variável resposta. No segundo cenário, os coeficientes da regressão foram $\beta_{01} =$

$0,5, \beta_{11} = 0,5, \beta_{21} = -0,5, \beta_{02} = 1,5, \beta_{12} = 0,5, \beta_{22} = -0,5, \beta_{03} = 0,2, \beta_{13} = -0,5, \beta_{23} = -0,5$, indicando que os melhores jogadores têm, em média probabilidades ligeiramente maiores de ganhar o primeiro *set*, e alta probabilidade de vitória em um *set* dado que venceu o anterior. Além disso, as covariáveis possuem um efeito fraco com a variável resposta. No terceiro cenário, os coeficientes foram $\beta_{01} = 1,2, \beta_{11} = 0,5, \beta_{21} = -0,5, \beta_{02} = 2,2, \beta_{12} = 0,5, \beta_{22} = -0,5, \beta_{03} = 1, \beta_{13} = -0,5, \beta_{23} = -0,5$, indicando que os melhores jogadores possuem, em média, maiores probabilidades de vitória no primeiro *set* e que as probabilidades de ambos os jogadores vencerem um *set* aumentam caso tenham vencido o *set* anterior, mantendo-se um efeito fraco das covariáveis sobre a variável resposta. No último cenário, os parâmetros foram especificados de modo a refletir que os melhores jogadores apresentam uma probabilidade moderadamente alta de vencer o primeiro *set* e, uma vez que tenham vencido o *set* anterior, possuem uma probabilidade elevada de vencer o próximo. Além disso, as variáveis explicativas exercem um efeito forte sobre a variável resposta. Os valores dos coeficientes especificados foram: $\beta_{01} = 1,2, \beta_{11} = 1,3, \beta_{21} = -1,2, \beta_{02} = 1,5, \beta_{12} = 1,4, \beta_{22} = -1,2, \beta_{03} = 1,5, \beta_{13} = 1,1, \beta_{23} = -1,2$.

Os resultados do estudo de simulação são apresentados na Tabela 4.1. Observa-se que os estimadores de máxima verossimilhança dos parâmetros do vetor $\boldsymbol{\beta}_1$ apresentam, em geral, vieses relativamente elevados para amostras pequenas, bem como valores de REQM também altos nesse caso. Nota-se, ainda, que os estimadores de máxima verossimilhança dos parâmetros dos vetores $\boldsymbol{\beta}_2$ e $\boldsymbol{\beta}_3$ exibem vieses pequenos e REQM moderados para pequenas amostras, com valores inferiores aos observados para o vetor $\boldsymbol{\beta}_1$. No entanto, conforme esperado, todos os vieses e REQM vão diminuindo a medida que aumentamos o tamanho da amostra. Além disso, para amostras a partir do tamanho 500, os vieses e os REQM são pequenos em todos os cenários.

Ao comparar os diferentes cenários, pode-se notar que os valores dos vieses e REQM no primeiro e segundo cenário são mais elevados em relação ao terceiro e quarto cenário, sendo que, para pequenas amostras, o quarto cenário, em geral, apresentou os menores valores. Para amostras grandes, os vieses e REQM dos estimadores não são muito diferentes para todos os cenários.

Tabela 4.1 – Estimativas do viés e REQM (em parênteses) dos estimadores do modelo considerando dependência entre os *sets* com os dados simulados.

Coeficiente	Valor	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
β_{01}	1,2	3,5834 (20,9442)	1,1074 (10,3695)	0,0417 (1,4079)	0,0334 (0,6346)	0,0186 (0,4517)
β_{11}	0,5	4,3748 (33,4726)	1,1786 (18,8128)	0,2173 (1,7824)	0,0435 (0,8968)	0,0188 (0,6424)
β_{21}	− 0,5	0,1902 (30,1357)	0,8995 (27,8810)	0,1466 (2,6896)	0,0171 (0,9284)	0,0142 (0,6393)
β_{02}	1,5	0,5430 (4,9716)	0,1467 (1,2022)	0,0795 (0,6938)	0,0268 (0,3845)	0,0106 (0,2730)
β_{12}	0,5	0,6555 (9,1606)	0,0249 (1,5357)	− 0,0064 (0,8963)	0,0172 (0,5379)	0,0094 (0,3832)
β_{22}	− 0,5	− 0,3370 (7,9774)	− 0,0183 (1,3249)	− 0,0369 (0,8878)	− 0,0145 (0,5447)	− 0,0039 (0,3818)
β_{03}	0,2	− 0,0185 (2,2646)	0,0336 (0,7892)	0,0110 (0,6142)	0,0062 (0,3341)	0,0069 (0,2384)
β_{13}	− 0,5	− 0,3502 (6,2823)	− 0,0285 (1,1352)	− 0,0044 (0,7919)	− 0,0055 (0,4477)	0,0007 (0,3190)
β_{23}	− 0,5	− 0,0199 (3,8921)	− 0,0168 (1,1231)	− 0,0127 (0,7579)	0,0114 (0,4482)	0,0018 (0,3181)
β_{01}	0,5	1,3529 (11,2847)	0,2528 (5,5574)	0,0182 (0,9444)	0,0254 (0,4934)	0,0086 (0,3497)
β_{11}	0,5	2,3933 (25,0993)	0,3213 (4,2449)	0,1122 (1,2682)	0,0193 (0,6870)	0,0182 (0,4830)
β_{21}	− 0,5	0,3192 (22,1954)	− 0,0140 (7,8727)	0,0800 (1,2634)	0,0195 (0,7117)	0,0090 (0,4769)
β_{02}	1,5	0,4535 (4,9966)	0,1316 (1,0037)	0,0641 (0,7372)	0,0224 (0,4067)	0,0090 (0,2899)
β_{12}	0,5	0,3169 (4,0746)	− 0,0169 (1,4558)	− 0,0034 (0,9524)	0,0151 (0,5631)	0,0031 (0,4028)
β_{22}	− 0,5	− 0,3927 (5,9413)	− 0,0462 (1,3435)	− 0,0637 (0,9442)	− 0,0353 (0,5684)	− 0,0078 (0,3939)
β_{03}	0,2	0,0028 (2,5947)	0,0375 (0,7193)	0,0100 (0,5537)	0,0120 (0,2998)	0,0057 (0,2135)
β_{13}	− 0,5	− 0,1997 (2,7947)	− 0,0210 (1,0296)	− 0,0008 (0,7130)	− 0,0093 (0,4053)	0,0023 (0,2873)
β_{23}	− 0,5	0,0291 (3,0050)	− 0,0241 (0,9789)	0,0098 (0,6887)	0,0035 (0,4063)	− 0,0014 (0,2830)
β_{01}	1,2	0,7342 (7,1524)	0,1173 (1,0636)	0,0353 (0,7207)	0,0207 (0,3866)	0,0026 (0,2766)
β_{11}	0,5	0,9645 (11,1156)	0,0842 (1,5978)	0,0275 (0,9544)	0,0181 (0,5309)	0,0099 (0,3804)
β_{21}	− 0,5	− 0,2470 (10,2650)	0,0118 (1,7825)	0,0142 (0,9404)	− 0,0131 (0,5486)	0,0036 (0,3704)
β_{02}	2,2	1,7970 (15,1165)	0,3702 (5,9608)	0,0739 (0,7679)	0,0270 (0,4096)	0,0174 (0,2897)
β_{12}	0,5	1,0145 (18,6461)	− 0,0449 (5,2464)	0,0390 (0,9826)	0,0137 (0,5561)	0,0066 (0,3957)
β_{22}	− 0,5	− 0,6753 (22,0752)	0,0559 (3,9295)	− 0,0234 (0,9731)	− 0,0093 (0,5588)	− 0,0080 (0,3822)
β_{03}	1,0	0,2810 (3,7194)	0,0808 (0,8696)	0,0222 (0,6397)	0,0148 (0,3486)	0,0074 (0,2486)
β_{13}	− 0,5	− 0,2769 (5,2030)	− 0,0704 (1,2045)	− 0,0178 (0,8018)	− 0,0092 (0,4501)	0,0008 (0,3210)
β_{23}	− 0,5	− 0,1665 (5,4620)	− 0,0289 (1,1386)	− 0,0019 (0,7744)	− 0,0042 (0,4511)	− 0,0071 (0,3223)
β_{01}	1,2	0,4715 (7,4293)	0,0765 (0,9035)	0,0312 (0,6645)	0,0166 (0,3642)	0,0063 (0,2591)
β_{11}	1,3	0,7378 (7,5066)	0,1245 (1,3165)	0,0502 (0,8781)	0,0170 (0,5072)	0,0032 (0,3576)
β_{21}	− 1,2	− 0,6793 (9,0732)	− 0,1253 (1,2480)	− 0,0444 (0,8445)	− 0,0235 (0,5102)	− 0,0069 (0,3527)
β_{02}	1,5	0,1809 (1,1258)	0,0751 (0,6952)	0,0281 (0,5231)	0,0159 (0,2848)	0,0068 (0,2016)
β_{12}	1,4	0,1904 (1,8392)	0,0246 (0,9859)	0,0196 (0,6742)	0,0082 (0,3838)	0,0074 (0,2752)
β_{22}	− 1,2	− 0,1342 (3,7433)	− 0,0301 (0,9474)	− 0,0185 (0,6535)	− 0,0145 (0,3909)	− 0,0046 (0,2670)
β_{03}	1,5	0,5062 (5,0656)	0,1300 (1,0326)	0,0327 (0,7071)	0,0195 (0,3848)	0,0172 (0,2736)
β_{13}	1,1	0,8524 (9,6670)	0,1535 (1,4578)	0,0889 (0,9242)	0,0080 (0,5000)	0,0020 (0,3703)
β_{23}	− 1,2	− 0,7335 (6,7530)	− 0,2109 (1,4683)	− 0,0476 (0,8963)	− 0,0101 (0,5165)	− 0,0158 (0,3596)

4.4 Aplicação em dados reais

Os torneios de Grand Slam constituem os eventos de maior prestígio no tênis profissional, oferecendo as maiores premiações e a maior pontuação no *ranking* mundial. Os quatro torneios que compõem o Grand Slam são: Australian Open, Roland Garros, Wimbledon e US Open, realizados nesta ordem ao longo do calendário competitivo. Cada torneio apresenta estrutura semelhante, com diversas categorias de competição organizadas em sistema eliminatório simples (perdeu, está eliminado). As principais categorias são: simples (masculino e feminino), duplas (masculino e feminino) e duplas mistas. As partidas da categoria simples masculina são disputadas no formato melhor de cinco *sets*, sendo vencedor o jogador que conquistar três *sets*. Já na categoria simples feminina, o formato é melhor de três *sets*, vencendo a jogadora que alcançar dois *sets*.

O conjunto de dados utilizado para aplicação do modelo de regressão proposto foi obtido a partir da primeira rodada da chave de simples masculina dos torneios disputados em 2024, totalizando 64 partidas por torneio e, sendo que houve duas disputas canceladas, portanto, um total de $n = 190$ observações. Os torneios considerados foram: Roland Garros, realizado em maio, em Paris (França); Wimbledon, realizado entre junho e julho, em Londres (Inglaterra); e o US Open, realizado entre agosto e setembro, em Nova Iorque (Estados Unidos). Todas as partidas foram disputadas no formato melhor de cinco *sets*, de modo que, neste caso, temos $k = 3$.

As variáveis selecionadas incluem os resultados das partidas da primeira rodada, além dos pontos e *rankings* dos jogadores, obtidos em datas imediatamente anteriores à realização dos respectivos torneios. Adicionalmente, foram incluídas duas variáveis complementares: *tourn played* (TP), que representa a contagem do número de torneios disputados pelo jogador durante os 12 meses anteriores; e *dropping*, que indica os pontos que o jogador perderá em breve, por estarem prestes a expirar, ou seja, a serem excluídos do *ranking* na data correspondente. Ambas as variáveis estão disponíveis no site oficial da ATP: <https://www.atptour.com/en/rankings/singles>.

A Tabela 4.2 apresenta as 6 primeiras linhas do banco de dados. Os dados apresentados nessa tabela foram reorganizados de forma que o jogador com melhor *ranking* seja sempre associado ao eixo x na análise bivariada (x_i, y_i). Essa organização visa facilitar tanto a análise estatística quanto a interpretação dos resultados.

Tabela 4.2 – Primeiras observações dos dados: Torneios de Grand Slam 2025.

ID	RANK_X	POINTS_X	TP_X	DROPPING_X	PLACAR_X	PLACAR_Y	DROPPING_Y	TP_Y	POINTS_Y	RANK_Y
1	1	9960	18	2000	3	0	2	22	450	142
2	63	816	32	45	3	1	35	32	669	91
3	40	1092	32	0	3	0	25	32	723	78
4	13	2700	23	360	3	0	21	29	831	62
5	53	940	30	35	3	1	35	15	492	135
6	12	2980	25	90	3	1	10	34	761	71

Para a análise e determinação das variáveis explicativas, algumas transformações nas variáveis originais são necessárias. Utilizaremos, principalmente, a diferença e a razão entre as variáveis observadas. Assim, definimos as seguintes variáveis:

- Drank = RANK_Y − RANK_X: diferença entre os *rankings*;
- Rrank = RANK_X/RANK_Y: razão entre os *rankings*;
- Dpts = POINTS_Y − POINTS_X: diferença entre os pontos;
- Rpts = POINTS_X/POINTS_Y: razão entre os pontos;
- Ddrop = DROPPING_Y − DROPPING_X: diferença entre os *droppings*;

- Rdrop = DROPPING_X/ DROPPING_Y: razão entre os *droppings*;
- Dtp = TP_Y − TP_X: diferença entre os *tourn played*;
- Rtp = TP_X/TP_Y: razão entre os *tourn played*.

Com o objetivo de comparar o desempenho do modelo de regressão proposto em (4.1), ajustamos o modelo independente descrito em (2.1) e o comparamos com o modelo dependente proposto, ambos utilizando a função de ligação logito nos parâmetros. Para seleção das covariáveis e identificação do modelo mais adequado, consideramos tanto os resultados de testes de Wald quanto o AIC. A Tabela 4.3 apresenta o melhor modelo ajustado sob a suposição de independência e o melhor modelo ajustado sob a suposição de dependência Markoviana, ambos avaliando o conjunto de covariáveis disponíveis. Observa-se que o modelo de regressão proposto apresenta menores valores dos critérios AIC e BIC, sendo, em princípio, mais adequado que o modelo que supõe independência para o conjunto de dados analisado.

Tabela 4.3 – Critérios de seleção de modelos.

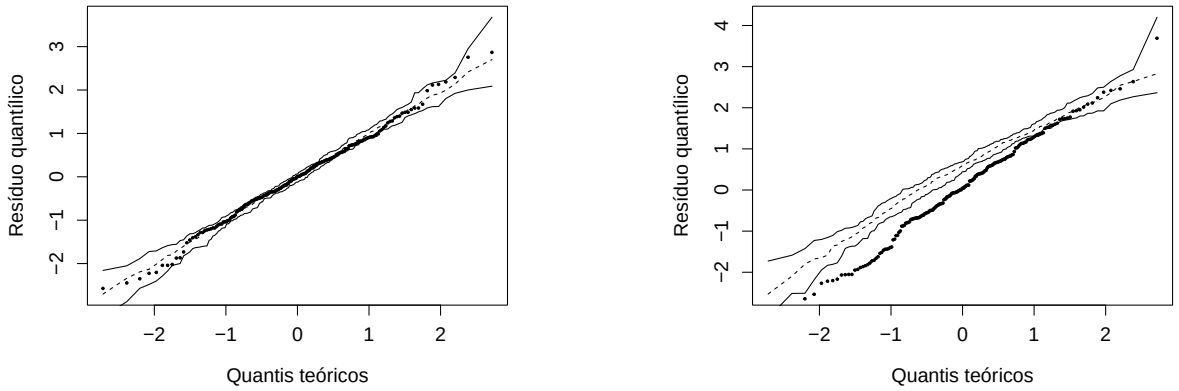
Modelos	AIC	BIC
Modelo independente : $g(p_i) = \beta_0 + \beta_1 Ddrop$	630,08	636,57
Modelo dependente:	609,69	629,17
$g_1(p_{0_i}) = \beta_0 + \beta_1 Dpts$		
$g_2(p_{00_i}) = \beta_0 + \beta_1 Ddrop$		
$g_3(p_{11_i}) = \beta_0 + \beta_1 Rpts$		

As estimativas dos parâmetros, os erros padrões, os valores observados da estatística de teste e os valores de p para o modelo considerado são apresentados na Tabela 4.4. Observam-se os sinais das estimativas dos parâmetros. Conclui-se que quanto maior é a diferença dos pontos dos jogadores, maior a probabilidade de o jogador de melhor *ranking* vencer o primeiro *set*. Da mesma forma, quanto maior a diferença entre os *droppings* dos jogadores maior a probabilidade do jogador de melhor *ranking* vencer um *set*, dado que venceu o *set* anterior. Observa-se também que quanto maior a razão entre os pontos dos jogadores, menor é a probabilidade do jogador B vencer um *set*, se ele venceu o *set* anterior. Note que esses três sinais das estimativas dos parâmetros estão de acordo com o que era esperado. Além disso, podemos interpretar as estimativas dos parâmetros. Estima-se por exemplo, que a cada aumento de uma unidade na diferença entre os pontos dos jogadores, aumenta em $0,07\%$ $((\exp(0,0007) - 1) \times 100\%)$ a chance do jogador de melhor *ranking* vencer o primeiro *set*. As demais estimativas dos parâmetros podem ser interpretadas de forma análoga.

Tabela 4.4 – Estimativas do modelo de regressão para os dado torneios de Grand Slam.

Submodelo	Covariável	Estimativa	Erro padrão	Estatística Z	Valor-p	Exp(Est.)
p_{0i}	Intercepto	$-0{,}3306$	0,3916	$-0{,}8442$	0,3985	0,7185
	Dpts	0,0007	0,0004	1,8087	0,0705	1,0007
p_{00i}	Intercepto	0,9520	0,2123	4,4850	$<0{,}0001$	2,5909
	Ddrop	0,0015	0,0009	1,7181	0,0858	1,0015
p_{11i}	Intercepto	0,4270	0,2738	1,5593	0,1189	1,5327
	Rpts	$-0{,}1393$	0,0764	$-1{,}8241$	0,0681	0,8700

Para a análise de diagnóstico, utilizamos o gráfico de probabilidade normal com envelope simulado para o resíduo quantílico aleatorizado, obtido a partir de w_i. A Figura 4.1 apresenta os envelopes para ambos os modelos, o que considerada a dependência entre os *sets* e o que considera independência entre os *sets*. Pela Figura 4.1 (a), vemos que o modelo proposto apresentou bom comportamento dos resíduos, com todos os pontos dentro do envelope, ou bem próximos a ele, indicando que o modelo proposto parece ajustar bem aos dados. Em contraste ao que é observado por meio do envelope do modelo que considerada independência entre os *sets*, conforme a Figura 4.1 (b), que indica um ajuste inadequado aos dados.



(a) Modelo dependente (b) Modelo independente

Figura 4.1 – Gráfico de probabilidade normal com envelope simulado dos modelos

A Figura 4.2 apresenta o gráfico de afastamento da verossimilhança para as 190 observações, utilizado na identificação de pontos potencialmente influentes. As observações $\{21\}$, $\{65\}$, $\{114\}$ e $\{131\}$ exibem valores mais elevados de deslocamento da verossimilhança. Para avaliar o impacto sobre as inferências do modelo proposto, realizamos o ajuste do modelo sem cada uma dessas observações potencialmente influentes, bem como considerando a exclusão em conjunto.

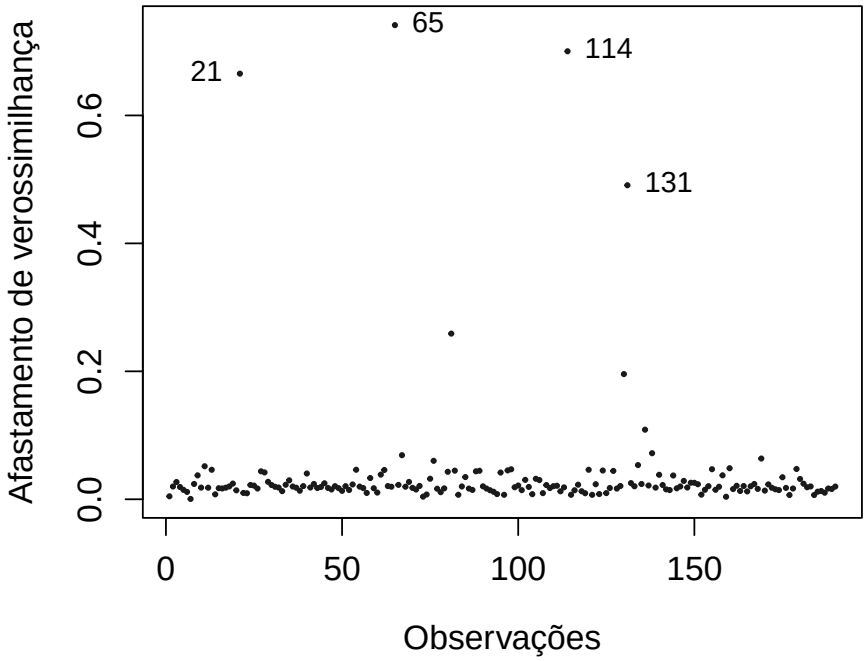


Figura 4.2 – Gráfico de afastamento de verossimilhança do modelo

Entre as medidas comumente utilizadas para quantificar o efeito da remoção de observações no ajuste do modelo, destacam-se a mudança relativa (RC), que quantifica a variação percentual na estimativa de um parâmetro após a remoção da observação, e a mudança relativa no erro-padrão (RCSE), que quantifica a variação percentual no erro-padrão da estimativa sob a mesma condição. Essas medidas são definidas como

$$\mathrm{RC}(\hat{\theta}_j)_{(i)} = \left| \frac{\hat{\theta}_j - \hat{\theta}_{j(i)}}{\hat{\theta}_j} \right| \times 100\% \quad \text{e} \quad \mathrm{RCSE}(\hat{\theta}_j)_{(i)} = \left| \frac{\mathrm{SE}(\hat{\theta}_j) - \mathrm{SE}(\hat{\theta}_j)_{(i)}}{\mathrm{SE}(\hat{\theta}_j)} \right| \times 100\%,$$

em que $\hat{\theta}_{j(i)}$ e $\mathrm{SE}(\hat{\theta}_j)_{(i)}$ representam, respectivamente, as estimativas de máxima verossimilhança do j-ésimo parâmetro do modelo e os erros padrões correspondentes, obtidos após a remoção da i-ésima observação.

A Tabela 4.5 apresenta as mudanças relativas nas estimativas dos parâmetros e as correspondentes mudanças nos erros padrões estimados. Observa-se que, em geral, essas medidas não apresentam valores elevados, exceto na exclusão da observação $\{65\}$, que causa uma alteração considerável no submodelo associado a p_{00_i}. Essa observação corresponde a uma partida com placar $(3,1)$, em que o jogador x_{65} (Jannik Sinner) ocupava a primeira posição no *ranking* mundial na época, enquanto o jogador y_{65} (Y. Hanfmann) encontrava-se na 110ª posição. Por esse motivo, as variáveis referentes às diferenças de pontos, de posições no *ranking* e de

droppings assumem valores elevados. Além disso, dada a diferença de nível dos jogadores, havia uma alta probabilidade estimada de que o jogo fosse vencido por Sinner por 3 sets a 0. Como isso não ocorreu, essa observação foi a que apresentou o maior valor de LD_i.

Embora os valores de RC e RCSE para o estimador do parâmetro associado à covariável presente no submodelo p_{00_i} se alterem consideravelmente com a exclusão da observação $\{65\}$, o valor de p sofre variação de apenas cerca de $0,02$ unidades. Além disso, a exclusão dessa observação não modifica de forma relevante as estimativas dos demais parâmetros do modelo. Assim, em termos práticos, o modelo ajustado com ou sem essa observação apresenta resultados muito semelhantes.

Tabela 4.5 – RCs (em %) nas estimativas e erros padrões do modelo correspondente para os casos removidos indicados e valores de p dos dados do torneios de Grand Slam.

Casos removidos	**Submodelo**	**Covariável**	**RC**$(\hat{\theta})$	**RCSE**$(\hat{\theta})$	**Valor-**p
$\{21\}$	p_{0i}	Intercepto	27,1078	0,8603	0,5418
		Dpts	10,9870	0,6644	0,1097
	p_{00i}	Intercepto	1,3617	1,4243	$<0,0001$
		Ddrop	5,8451	10,1816	0,0988
	p_{11i}	Intercepto	25,1173	10,7648	0,0782
		Rpts	38,8446	25,7920	0,0441
$\{114\}$	p_{0i}	Intercepto	31,2686	5,6361	0,2941
		Dpts	39,6919	24,9673	0,0432
	p_{00i}	Intercepto	4,0034	3,1540	$<0,0001$
		Ddrop	7,2803	14,3168	0,0630
	p_{11i}	Intercepto	1,4616	0,5553	0,1265
		Rpts	2,7055	1,8701	0,0659
$\{65\}$	p_{0i}	Intercepto	22,5727	5,7662	0,4879
		Dpts	24,2670	25,8852	0,0646
	p_{00i}	Intercepto	1,4522	1,8968	$<0,0001$
		Ddrop	66,8388	78,1591	0,1076
	p_{11i}	Intercepto	2,0322	0,7348	0,1294
		Rpts	0,6039	0,9365	0,0724
$\{131\}$	p_{0i}	Intercepto	8,8476	0,8016	0,3620
		Dpts	8,2874	2,5150	0,0561
	p_{00i}	Intercepto	1,7261	0,9085	$<0,0001$
		Ddrop	35,0142	25,7191	0,0650
	p_{11i}	Intercepto	4,0383	0,8333	0,1313
		Rpts	7,3432	1,9525	0,0847
$\{21,114,65,131\}$	p_{0i}	Intercepto	11,2369	6,0419	0,3758
		Dpts	29,5520	22,8707	0,0565
	p_{00i}	Intercepto	7,4857	1,8282	$<0,0001$
		Ddrop	72,8340	46,3324	0,0424
	p_{11i}	Intercepto	14,9494	10,6159	0,1051
		Rpts	30,5807	27,1150	0,0609

Um comparativo que avalia o desempenho do modelo proposto é apresentado na Tabela

4.6, que traz para o modelo proposto e o modelo independente, a estimativa do número de jogos com cada um dos possíveis placares. Observa-se que o modelo dependente apresenta previsões mais próximas dos placares reais observados, especialmente nos resultados mais desequilibrados (3×0, 0×3). Por sua vez, o modelo independente tende a superestimar a quantidades de jogos equilibrados (como 3×2) e subestimar a quantidade de jogos com placares desequilibrados (como 3×0, 0×3), uma vez que assume que a probabilidade de um jogador vencer um *set* é independente dos demais *sets*. Portanto, o modelo proposto com dependência Markoviana entre os *sets* apresentou melhor desempenho e precisão no conjunto de dados reais, enquanto o modelo que considera independência entre os *sets* é considerado inadequado para representar a estrutura real dos resultados.

Tabela 4.6 – Distribuição de frequências dos resultados dos placares e previsões dos modelos.

(x,y)	Freq. absoluta	Freq. relativa	Freq. absoluta prevista	
			Modelo dependente	Modelo independente
(3,0)	64	0,34	64,86	52,85
(3,1)	47	0,25	42,34	50,05
(3,2)	23	0,12	28,85	36,37
(0,3)	21	0,11	20,63	10,06
(1,3)	15	0,08	16,48	18,32
(2,3)	20	0,10	16,84	22,35
Total	190	1	190	190

CAPÍTULO

5

CONCLUSÕES

O objetivo desta tese foi propor distribuições de probabilidade e modelos de regressão para placares de jogos esportivos divididos em *sets*. A principal motivação para esses modelos, além de contribuir com o desenvolvimento de novas distribuições de probabilidade e de novos modelos de regressão, foi propor uma forma alternativa e adequada de analisar as probabilidades de vitória de um jogo, modelando diretamente os placares. Consideramos tanto o caso em que há independência entre os *sets*, como a situação em que há dependência Markoviana entre os mesmos.

Com base nos modelos de regressão propostos, foram discutidas ferramentas inferenciais, tais como expressões em forma fechada para a função escore e para a matriz de informação de Fisher. Também foram abordadas a estimação dos parâmetros pelo método da máxima verossimilhança e a obtenção de estimativas intervalares e de testes de hipóteses. Adicionalmente, foram apresentadas medidas de diagnóstico voltadas a avaliação da adequação do modelo a um conjunto de dados. Os parâmetros dos modelos podem variar em função de variáveis explicativas e são interpretáveis se usamos a função de ligação logito. Os estudos de simulação de Monte Carlo sugerem que o estimador de máxima verossimilhança do modelo tem boas propriedades.

Para finalizar o trabalho, apresentamos aplicações dos modelos propostos em dados reais de tênis e tênis de mesa. Na aplicação que utilizou dados dos torneios de Grand Slam de tênis profissional, discutimos a validade e a utilidade da modelagem proposta, demonstrando que o modelo pode ser empregado para a previsão dos placares de jogos esportivos divididos em *sets*. Essa abordagem mostra-se particularmente útil para o planejamento dos jogos de um determinado dia em um torneio, considerando diferentes quadras do local dos jogos. Adicionalmente, o trabalho abre caminho para futuros estudos aplicados que poderão considerar um número ainda maior de covariáveis no modelo.

5.0.1 Trabalhos futuros

Os trabalhos futuros que podem ser desenvolvidos com base nesta tese incluem:

- Desenvolver modelos de regressão para os placares de resultados de jogos esportivos divididos em *sets*, considerando outras estruturas de dependência entre os *sets* que não a Markoviana.
- Desenvolver correções analíticas e por *bootstrap* para estatísticas de testes que podem ser utilizadas no modelo, como a estatística da razão de verossimilhanças. Usualmente se utiliza para essa estatística a correção de Bartlett (BARTLETT, 1937). Há garantias teóricas que essa correção melhora a estatística da razão de verossimilhanças apenas em modelos de regressão com respostas contínuas. No entanto, na prática, ela muitas vezes melhora a estatística também quando a resposta é discreta (MOULTON; WEISSFELD; LAURENT, 1993; DAS; DHAR; PRADHAN, 2018) e quando a resposta tem parte discreta e parte contínua (MAGALHÃES *et al.*, 2024).

REFERÊNCIAS

AKI, S.; HIRANO, K. Discrete distributions related to succession events in a two-state markov chain. **Statistical science and data analysis**, VSP International Science Publishers Zeist, p. 467–474, 1993. Citado na página 20.

ATKINSON, A. C. Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. Oxford Science Publications, 1985. Citado na página 32.

BARNETT, T.; CLARKE, S. R. Combining player statistics to predict outcomes of tennis matches. **IMA Journal of Management Mathematics**, Oxford University Press, v. 16, n. 2, p. 113–120, 2005. Citado na página 20.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the royal society of london. series a-mathematical and physical sciences**, The Royal Society London, v. 160, n. 901, p. 268–282, 1937. Citado na página 66.

BAZÁN, J.; TORRES-AVILÉS, F.; SUZUKI, A. K.; LOUZADA, F. Power and reversal power links for binary regressions: An application for motor insurance policyholders. **Applied Stochastic Models in Business and Industry**, Wiley Online Library, v. 33, n. 1, p. 22–34, 2017. Citado na página 28.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. [S.l.]: SBM, 2001. v. 2. Citado na página 27.

CARRARI, A.; FERRANTE, M.; FONSECA, G. *et al.* A new markovian model for tennis matches. **Electronic Journal of Applied Statistical Analysis**, v. 10, n. 3, p. 693–711, 2017. Citado na página 19.

COOK, R. D.; PEÑA, D.; WEISBERG, S. **Residuals and influence in regression**. [S.l.]: Chapman-Hall, 1982. Citado na página 56.

CORTÉS, I. E.; CASTRO, M. de; GALLARDO, D. I. A new family of quantile regression models applied to nutritional data. **Journal of Applied Statistics**, Taylor & Francis, p. 1–21, 2023. Citado na página 56.

DAS, U.; DHAR, S. S.; PRADHAN, V. Corrected likelihood-ratio tests in logistic regression using small-sample data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 47, n. 17, p. 4272–4285, 2018. Citado na página 66.

DIMITROV, B. N.; KOLEV, N. Bernoulli trials: Extensions, related probability distributions and modeling powers. 2002. Citado na página 20.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 31.

FABIO, L. C.; VILLEGAS, C.; CARRASCO, J. M.; CASTRO, M. d. Diagnostic tools for a multivariate negative binomial model for fitting correlated data with overdispersion. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 52, n. 6, p. 1833–1853, 2023. Citado na página 56.

KOLEV, N.; MINKOVA, L.; NEYTCHEV, P. Inflated-parameter family of generalized power series distributions and their application in analysis of overdispersed insurance data. **ARCH Research Clearing House**, v. 2, p. 295–320, 2000. Citado na página 20.

LEHMANN, E. L.; ROMANO, J. P.; CASELLA, G. **Testing statistical hypotheses**. [S.l.]: Springer, 1986. v. 3. Citado na página 31.

LEI, Y.; LIN, A.; CAO, J. Rhythms of victory: Predicting professional tennis matches using machine learning. **IEEE Access**, IEEE, 2024. Citado na página 20.

LEMONTE, A. **The gradient test: Another likelihood-based test**. [S.l.]: Academic Press, 2016. Citado na página 31.

LEMONTE, A. J.; BAZÁN, J. L. New links for binary regression: an application to coca cultivation in peru. **Test**, Springer, v. 27, p. 597–617, 2018. Citado na página 28.

LEWIS, M. **Moneyball: The art of winning an unfair game**. [S.l.]: WW Norton & Company, 2004. Citado na página 19.

LI, B.; DENG, Z.; GUPTA, G.; LI, J.; MIAO, Y. Predicting tennis match outcomes mid-game using machine learning on psychological and physical data. **Journal of Big Data**, Springer, v. 12, n. 1, p. 159, 2025. Citado na página 20.

MADURSKA, A. M. A set-by-set analysis method for predicting the outcome of professional singles tennis matches. **4th year Software Engineering MEng project, Imperial College London, Department of Computing: London, UK**, 2012. Citado na página 20.

MAGALHÃES, T. M.; PEREIRA, G. H.; BOTTER, D. A.; SANDOVAL, M. C. Bartlett corrections for zero-adjusted generalized linear models. **Statistical Papers**, Springer, v. 65, n. 4, p. 2191–2209, 2024. Citado na página 66.

MCCULLAGH, P.; NELDER, J. A. Generalized linear models 2nd edition chapman and hall. **London, UK**, 1989. Citado na página 28.

MILLAR, R. B. **Maximum likelihood estimation and inference: with examples in R, SAS and ADMB**. [S.l.]: John Wiley & Sons, 2011. Citado na página 30.

MOULTON, L. H.; WEISSFELD, L. A.; LAURENT, R. T. S. Bartlett correction factors in logistic regression models. **Computational statistics & data analysis**, Elsevier, v. 15, n. 1, p. 1–11, 1993. Citado na página 66.

NEWTON, P. K.; KELLER, J. B. Probability of winning at tennis i. theory and data. **Studies in applied Mathematics**, Wiley Online Library, v. 114, n. 3, p. 241–269, 2005. Citado na página 19.

O'MALLEY, A. J. Probability formulas and statistical analysis in tennis. **Journal of Quantitative Analysis in Sports**, De Gruyter, v. 4, n. 2, 2008. Citado na página 19.

PAULA, G. A. Modelos de regressao e aplicacoes. IME-USP São Paulo, 2025. Citado na página 31.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>. Citado na página 32.

SHUKLA, A.; KUMAR, G.; YADAV, T. Logistic regression model in success and failure of women's volleyball. **International Research Journal of Management Sociology Humanity**, 2020. Citado na página 20.

SIM, M. K.; CHOI, D. G. The winning probability of a game and the importance of points in tennis matches. **Research Quarterly for Exercise and Sport**, Taylor & Francis, v. 91, n. 3, p. 361–372, 2020. Citado na página 19.

SOUZA, R. d. Modelo geométrico de ordem k correlacionado. Universidade Federal de São Carlos, 2019. Citado na página 20.

SUZUKI, A. K. Modelagem estatística para a determinação de resultados de dados esportivos. Universidade Federal de São Carlos, 2007. Citado na página 19.

THORN, J.; PALMER, P. **The sabermetric revolution: Assessing the growth of analytics in baseball**. [S.l.]: Rowman Littlefield, 2009. Citado na página 19.

WANG, F.; WANG, Y.; LIAO, L. *et al.* Research and analysis of winning factors in men's singles tennis based on logistic regression modelling. **Frontiers in Sport Research**, Francis Academic Press, v. 6, n. 4, 2024. Citado na página 20.

YANG, D.; SMALL, D. S. An r package and a study of methods for computing empirical likelihood. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 83, n. 7, p. 1363–1372, 2013. Citado na página 29.

ZHAO, P.; LUO, T.; BI, P. Athlete performance analysis: Machine learning for predicting tennis player scores. In: IEEE. **2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)**. [S.l.], 2024. p. 1158–1162. Citado na página 20.

APÊNDICE

A

PROVAS E RESULTADOS

Neste apêndice são apresentadas as provas de alguns resultados desenvolvidos do Capítulo 2 e 4.

A.1 Cálculo do valor esperado da distribuição marginal

O objetivo é desenvolver a média da função de probabilidade marginal definida em (2.1), fixando $k = 4$. Pela expressão (2.4) o valor esperado μ_X é dado por

$$
\begin{aligned}
\mu_X &= \sum_x xP(X=x) = \sum_{i=0}^{k-1} iP(X=i,Y=k) + k\sum_{j=0}^{k-1} P(X=k,Y=j) = \\
&= 0\times P(X=0,Y=4) + 1\times P(X=1,Y=4) + 2\times P(X=2,Y=4) + 3\times P(X=3,Y=4) + \\
&\quad 4\times\{P(X=4,Y=0) + P(X=4,Y=1) + P(X=4,Y=2) + P(X=4,Y=3)\} \\
&= \binom{1+4-1}{4-1}p(1-p)^4 + 2\times\binom{2+4-1}{4-1}p^2(1-p)^4 + 3\times\binom{3+4-1}{4-1}p^3(1-p)^4 + \\
&\quad 4\times\left\{\binom{4+0-1}{4-1}p^4(1-p)^0 + \binom{4+1-1}{4-1}p^4(1-p)^1 + \binom{4+2-1}{4-1}p^4(1-p)^2 + \right. \\
&\quad \left.\binom{4+3-1}{4-1}p^4(1-p)^3\right\} \\
&= \binom{4}{3}p(1-p)^4 + 2\times\binom{5}{3}p^2(1-p)^4 + 3\times\binom{6}{3}p^3(1-p)^4 + \\
&\quad 4\times\left\{\binom{3}{3}p^4 + \binom{4}{3}p^4(1-p) + \binom{5}{3}p^4(1-p)^2 + \binom{6}{3}p^4(1-p)^3\right\} \\
&= 4p(1-p)^4 + 20p^2(1-p)^4 + 60p^3(1-p)^4 + \\
&\quad 4p^4 + 16p^4(1-p) + 40p^4(1-p)^2 + 80p^4(1-p)^3.
\end{aligned}
$$

Da mesma forma, podemos obter o valor esperado μ_Y, dado por

$$\begin{aligned}
\mu_Y &= E(Y) = \sum_y yP(Y=y) = \sum_{i=0}^{k-1} iP(X=k,Y=i) + k\sum_{j=0}^{k-1} P(X=j,Y=k) \\
&= 0\times P(X=4,Y=0) + 1\times P(X=4,Y=1) + 2\times P(X=4,Y=2) + 3\times P(X=4,Y=3) + \\
&\quad 4\times\{P(X=0,Y=4) + P(X=1,Y=4) + P(X=2,Y=4) + P(X=3,Y=4)\} \\
&= 1\times\binom{4+1-1}{4-1}p^4(1-p) + 2\times\binom{4+2-1}{4-1}p^4(1-p)^2 + 3\times\binom{4+3-1}{4-1}p^4(1-p)^3 + \\
&\quad 4\times\left\{\binom{0+4-1}{4-1}p^0(1-p)^4 + \binom{1+4-1}{4-1}p(1-p)^4 + \binom{2+4-1}{4-1}p^2(1-p)^4 + \right. \\
&\quad \left.\binom{3+4-1}{4-1}p^3(1-p)^4\right\} \\
&= \binom{4}{3}p^4(1-p) + 2\times\binom{5}{3}p^4(1-p)^2 + 3\times\binom{6}{3}p^4(1-p)^3 + \\
&\quad 4\times\left\{\binom{3}{3}(1-p)^4 + \binom{4}{3}p(1-p)^4 + \binom{5}{3}p^2(1-p)^4 + \binom{6}{3}p^3(1-p)^4\right\} \\
&= 4p^4(1-p) + 20p^4(1-p)^2 + 60p^4(1-p)^3 + \\
&\quad 4(1-p)^4 + 16p(1-p)^4 + 40p^2(1-p)^4 + 80p^3(1-p)^4.
\end{aligned}$$

A.2 Cálculo do vetor escore e da matriz de informação de Fisher

Nesta seção são apresentadas as provas de alguns resultados apresentados no Capítulo 4, referentes ao cálculo da função escore, e da matriz de informação de Fisher do modelo proposto.

A.2.1 Obtenção da função escore para o modelo

Derivadas de primeira ordem das probabilidades a_{ij} (3.1) com relação a p_{0i}.

$$\begin{aligned}
b_{31i} &= \left\{2p_{00i}(1-p_{00i})(1-p_{11i}) - p_{00i}^2(1-p_{11i})\right\} \\
b_{32i} &= \left\{2p_{00i}p_{11i}(1-p_{00i})(1-p_{11i}) + (1-p_{00i})^2(1-p_{11i})^2 - \right. \\
&\quad \left.[p_{00i}^2p_{11i}(1-p_{11i}) + 2p_{00i}(1-p_{00i})(1-p_{11i})^2]\right\} \\
b_{13i} &= \left\{p_{11i}^2(1-p_{00i}) - 2p_{11i}(1-p_{00i})(1-p_{11i})\right\} \\
b_{23i} &= \left\{\left[2p_{11i}(1-p_{00i})^2(1-p_{11i}) + p_{00i}p_{11i}^2(1-p_{00i})\right] - \right. \\
&\quad \left.\left[2p_{00i}p_{11i}(1-p_{00i})(1-p_{11i}) + (1-p_{00i})^2(1-p_{11i})^2\right]\right\}.
\end{aligned}$$

Derivadas de primeira ordem das probabilidades a_{ij} (3.1) com relação a p_{00i}.

$$\begin{aligned}
c_{31i} &= \{p_{0i}[2(1-p_{11i})(1-2p_{00i})]+p_{1i}[2p_{00i}(1-p_{11i})]\}\\
c_{32i} &= \{p_{0i}\left[2p_{11i}(1-p_{11i})(1-2p_{00i})+2(1-p_{11i})^2(p_{00i}-1)\right]+\\
&\quad p_{1i}\{\left[2(1-p_{11i})^2(1-2p_{00i})\right]+2\left[p_{00i}p_{11i}(1-p_{11i})\right]\}\}\\
c_{13i} &= \{-p_{0i}p_{11i}^2-p_{1i}2p_{11i}(1-p_{11i})\}\\
c_{23i} &= \{p_{0i}\left[4p_{11i}(1-p_{11i})(p_{00i}-1)+p_{11i}^2(1-2p_{00i})\right]+\\
&\quad p_{1i}\left[2p_{11i}(1-p_{11i})(1-2p_{00i})+2(1-p_{11i})^2(p_{00i}-1)\right]\}.
\end{aligned}$$

Derivadas de primeira ordem das probabilidades a_{ij} (3.1) com relação a p_{11i}.

$$\begin{aligned}
d_{31i} &= \{p_{0i}[-2p_{00i}(1-p_{00i})]-p_{1i}p_{00i}^2\}\\
d_{32i} &= \{p_{0i}\left[2p_{00i}(1-p_{00i})(1-2p_{11i})+2(1-p_{00i})^2(p_{11i}-1)\right]+\\
&\quad p_{1i}\left[4p_{00i}(1-p_{00i})(p_{11i}-1)+p_{00i}^2(1-2p_{11i})\right]\}\\
d_{13i} &= \{p_{0i}[2p_{11i}(1-p_{00i})]+p_{1i}[2(1-p_{00i})(1-2p_{11i})]\}\\
d_{23i} &= \{p_{0i}\left[2(1-2p_{11i})(1-p_{00i})^2+2p_{00i}p_{11i}(1-p_{00i})\right]+\\
&\quad p_{1i}\left[2p_{00i}(1-p_{00i})(1-2p_{11i})+2(p_{11i}-1)(1-p_{00i})^2\right]\}.
\end{aligned}$$

O componente do vetor escore (4.3) associado à β_{j1}, parâmetro de posição j do vetor $\boldsymbol{\beta}_1$, é obtido derivando-se o logaritmo da função de log-verssimilhança (4.2) em relação à β_{j1}. Logo, o componente é dado por

$$U_{\beta_{j1}}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\frac{\partial\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}}\frac{\partial p_{0i}}{\partial\eta_{i1}}\frac{\partial\eta_{i1}}{\partial\beta_{j1}} = \sum_{i=1}^{n}d_{p_{0i}}l_{p_{0i}}z_{ij1} \tag{A.1}$$

e

$$U_{\boldsymbol{\beta}_1}(\boldsymbol{\theta}) = \left(\frac{\partial\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial\beta_{01}},\ldots,\frac{\partial\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial\beta_{J1}}\right)^{\top},$$

em que

$$\begin{aligned}
d_{p_{0i}} &= \frac{\partial\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}} &\text{(A.2)}\\
&= \frac{I_{\{3,0\}}(x_i,y_i)}{p_{0i}}+I_{\{3,1\}}(x_i,y_i)\frac{b_{31i}}{a_{31i}}+I_{\{3,2\}}(x_i,y_i)\frac{b_{32i}}{a_{32i}}\\
&\quad -\frac{I_{\{0,3\}}(x_i,y_i)}{p_{1i}}+I_{\{1,3\}}(x_i,y_i)\frac{b_{13i}}{a_{13i}}+I_{\{2,3\}}(x_i,y_i)\frac{b_{23i}}{a_{23i}} &\text{(A.3)}
\end{aligned}$$

e

$$\begin{aligned}
l_{p_{0i}} &= \frac{\partial p_{0i}}{\partial\eta_{i1}} = \frac{1}{g_1'(p_{0i})}\\
l_{p_{00i}} &= \frac{\partial p_{00i}}{\partial\eta_{i2}} = \frac{1}{g_2'(p_{00i})}\\
l_{p_{11i}} &= \frac{\partial p_{11i}}{\partial\eta_{i3}} = \frac{1}{g_3'(p_{11i})},
\end{aligned}$$

que pode ser escrito na forma matricial como na equação (4.3). De forma equivalente,

$$U_{\beta_{j2}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}} \frac{\partial p_{00i}}{\partial \eta_{i2}} \frac{\partial \eta_{i2}}{\partial \beta_{j2}} = \sum_{i=1}^{n} d_{p_{00i}} l_{p_{00i}} z_{ij2}, \tag{A.4}$$

$$U_{\beta_{j3}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}} \frac{\partial p_{11i}}{\partial \eta_{i3}} \frac{\partial \eta_{i3}}{\partial \beta_{j3}} = \sum_{i=1}^{n} d_{p_{11i}} l_{p_{11i}} z_{ij3}, \tag{A.5}$$

em que

$$U_{\boldsymbol{\beta}_2}(\boldsymbol{\theta}) = \left(\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{02}}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{J2}} \right)^{\top},$$

$$U_{\boldsymbol{\beta}_3}(\boldsymbol{\theta}) = \left(\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{03}}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{J3}} \right)^{\top},$$

com

$$\begin{aligned} d_{p_{00i}} &= \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}} \\ &= I_{\{3,0\}}(x_i,y_i)\frac{2}{p_{00i}} + I_{\{3,1\}}(x_i,y_i)\frac{c_{31i}}{a_{31i}} + I_{\{3,2\}}(x_i,y_i)\frac{c_{32i}}{a_{32i}} \\ &\quad + I_{\{1,3\}}(x_i,y_i)\frac{c_{13i}}{a_{13i}} + I_{\{2,3\}}(x_i,y_i)\frac{c_{23i}}{a_{23i}}, \end{aligned}$$

$$\begin{aligned} d_{p_{11i}} &= \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}} \\ &= I_{\{3,1\}}(x_i,y_i)\frac{d_{31i}}{a_{31i}} + I_{\{3,2\}}(x_i,y_i)\frac{d_{32i}}{a_{32i}} + I_{\{0,3\}}(x_i,y_i)\frac{2}{p_{11i}} \\ &\quad + I_{\{1,3\}}(x_i,y_i)\frac{d_{13i}}{a_{13i}} + I_{\{2,3\}}(x_i,y_i)\frac{d_{23i}}{a_{23i}}, \end{aligned}$$

podem ser escritos na forma matricial, são apresentados em (4.3).

A.2.2 Obtenção da matriz de informação de Fisher

Para obter a matriz de informação de Fisher, calculamos o valor esperado da derivada de segunda ordem da função de log-verossimilhança (4.2) em relação a cada um dos parâmetros. Utilizando-se as derivadas do vetor escore já apresentadas em (4.3), temos que

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j1} \partial \beta_{k1}} &= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial p_{0i}} \left(\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}} \frac{\partial p_{0i}}{\partial \eta_{i1}} \frac{\partial \eta_{i1}}{\partial \beta_{j1}} \right) \frac{\partial p_{0i}}{\partial \eta_{i1}} \frac{\partial \eta_{i1}}{\partial \beta_{k1}} \right\} \\ &= \sum_{i=1}^{n} \left\{ \left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}^2} \left(\frac{\partial p_{0i}}{\partial \eta_{i1}} \right)^2 \frac{\partial \eta_{i1}}{\partial \beta_{j1}} \frac{\partial \eta_{i1}}{\partial \beta_{k1}} \right] + \right. \\ &\quad \left. \left[\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}} \left(\frac{\partial}{\partial p_{0i}} \frac{\partial p_{0i}}{\partial \eta_{i1}} \right) \frac{\partial p_{0i}}{\partial \eta_{i1}} \frac{\partial \eta_{i1}}{\partial \beta_{j1}} \frac{\partial \eta_{i1}}{\partial \beta_{k1}} \right] \right\} \\ &= \sum_{i=1}^{n} \left\{ \left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}^2} \left(\frac{\partial p_{0i}}{\partial \eta_{i1}} \right)^2 + \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})}{\partial p_{0i}} \left(\frac{\partial^2 p_{0i}}{\partial \eta_{i1}^2} \right) \right] z_{ij1} z_{ik1} \right\}. \end{aligned} \tag{A.6}$$

Temos que

$$\frac{\partial \eta_{i1}}{\partial \beta_{j1}} = z_{ij1} \qquad \text{e} \qquad \frac{\partial \eta_{i1}}{\partial \beta_{k1}} = z_{ik1}.$$

Sob condições de regularidade, $E\left(\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})/\partial p_{0i}\right) = 0$. Portanto, obtém-se que o componente (j,k) de $J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}$ é dado por

$$\begin{aligned} -E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j1}\partial \beta_{k1}}\right) &= \sum_{i=1}^{n}\left\{\left[-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}^2}\right]\left(\frac{\partial p_{0i}}{\partial \eta_{i1}}\right)^2\right] z_{ij1}z_{ik1}\right\} \\ &= \sum_{i=1}^{n} f_{p_{0i}}(l_{p_{0i}})^2 z_{ij1}z_{ik1}, \end{aligned}$$

em que

$$\begin{aligned} f_{p_{0i}} &= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}^2}\right] \\ &= \frac{p_{00i}^2}{p_{0i}} + \frac{(b_{31i})^2}{a_{31i}} + \frac{(b_{32i})^2}{a_{32i}} + \frac{p_{11i}^2}{p_{1i}} + \frac{(b_{13i})^2}{a_{13i}} + \frac{(b_{23i})^2}{a_{23i}}, \end{aligned}$$

sendo que $l_{p_{0i}}$ e b_{ij} foram definidos em A.2. Logo, podemos escrever $J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}$ na forma matricial conforme apresentado em (4.4).

De forma equivalente,

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j2}\partial \beta_{k2}} &= \sum_{i=1}^{n}\left\{\frac{\partial}{\partial p_{00i}}\left(\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}}\frac{\partial p_{00i}}{\partial \eta_{i2}}\frac{\partial \eta_{i1}}{\partial \beta_{j2}}\right)\frac{\partial p_{00i}}{\partial \eta_{i2}}\frac{\partial \eta_{i2}}{\partial \beta_{k2}}\right\} \\ &= \sum_{i=1}^{n}\left\{\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}^2}\left(\frac{\partial p_{00i}}{\partial \eta_{i2}}\right)^2\frac{\partial \eta_{i2}}{\partial \beta_{j2}}\frac{\partial \eta_{i2}}{\partial \beta_{k2}}\right] + \right. \\ &\quad \left.\left[\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}}\left(\frac{\partial}{\partial p_{00i}}\frac{\partial p_{00i}}{\partial \eta_{i2}}\right)\frac{\partial p_{00i}}{\partial \eta_{i2}}\frac{\partial \eta_{i2}}{\partial \beta_{j2}}\frac{\partial \eta_{i2}}{\partial \beta_{k2}}\right]\right\} \\ &= \sum_{i=1}^{n}\left\{\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}^2}\left(\frac{\partial p_{00i}}{\partial \eta_{i2}}\right)^2 + \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})}{\partial p_{00i}}\left(\frac{\partial^2 p_{00i}}{\partial \eta_{i2}^2}\right)\right] z_{ij2}z_{ik2}\right\}, \end{aligned}$$

em que

$$\frac{\partial \eta_{i2}}{\partial \beta_{j2}} = z_{ij2} \qquad \text{e} \qquad \frac{\partial \eta_{i2}}{\partial \beta_{k2}} = z_{ik2}.$$

Sob condições de regularidade, $E\left(\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})/\partial p_{00i}\right) = 0$. Portanto, obtém-se que o componente (j,k) de $J_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}$ é dado por

$$\begin{aligned} -E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j2}\partial \beta_{k2}}\right) &= \sum_{i=1}^{n}\left\{\left[-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}^2}\right]\left(\frac{\partial p_{00i}}{\partial \eta_{i2}}\right)^2\right] z_{ij2}z_{ik2}\right\} \\ &= \sum_{i=1}^{n} f_{p_{00i}}(l_{p_{00i}})^2 z_{ij2}z_{ik2}, \end{aligned}$$

em que

$$\begin{aligned} f_{p_{00i}} &= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}^2}\right] \\ &= 2p_{0i} - \frac{h_{31i}a_{31i} - (c_{31i})^2}{a_{31i}} - \frac{h_{32i}a_{32i} - (c_{32i})^2}{a_{32i}} \\ &\quad + \frac{(c_{13i})^2}{a_{13i}} - \frac{h_{23i}a_{23i} - (c_{23i})^2}{a_{23i}}, \end{aligned}$$

em que h_{ij} representa a derivada de segunda ordem das probabilidades a_{ij} com relação a p_{00i}

$$\begin{aligned} h_{31i} &= \{p_{0i}[-4(1-p_{11i})] + p_{1i}[2(1-p_{11i})]\} \\ h_{32i} &= \left\{p_{0i}\left[2(1-p_{11i})^2 - 4p_{11i}(1-p_{11i})\right] + \right. \\ &\quad \left. p_{1i}\left[2p_{11i}(1-p_{11i}) - 4(1-p_{11i})^2\right]\right\} \\ h_{13i} &= 0 \\ h_{23i} &= p_{0i}\left[4p_{11i}(1-p_{11i}) - 2p_{11i}^2\right] + p_{1i}\left[-4p_{11i}(1-p_{11i}) + 2(1-p_{11i})^2\right]. \end{aligned}$$

Então, escrevemos $K_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}$ na forma matricial, conforme apresentado em (4.4).

De forma equivalente,

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j3}\partial \beta_{k3}} &= \sum_{i=1}^{n}\left\{\frac{\partial}{\partial p_{11i}}\left(\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}}\frac{\partial p_{11i}}{\partial \eta_{i3}}\frac{\partial \eta_{i3}}{\partial \beta_{j3}}\right)\frac{\partial p_{11i}}{\partial \eta_{i3}}\frac{\partial \eta_{i3}}{\partial \beta_{k3}}\right\} \\ &= \sum_{i=1}^{n}\left\{\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}^2}\left(\frac{\partial p_{11i}}{\partial \eta_{i3}}\right)^2\frac{\partial \eta_{i3}}{\partial \beta_{j3}}\frac{\partial \eta_{i3}}{\partial \beta_{k3}}\right] + \right. \\ &\quad \left.\left[\frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}}\left(\frac{\partial}{\partial p_{11i}}\frac{\partial p_{11i}}{\partial \eta_{i3}}\right)\frac{\partial p_{11i}}{\partial \eta_{i3}}\frac{\partial \eta_{i3}}{\partial \beta_{j3}}\frac{\partial \eta_{i3}}{\partial \beta_{k3}}\right]\right\} \\ &= \sum_{i=1}^{n}\left\{\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}^2}\left(\frac{\partial p_{11i}}{\partial \eta_{i3}}\right)^2 + \frac{\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})}{\partial p_{11i}}\left(\frac{\partial^2 p_{11i}}{\partial \eta_{i3}^2}\right)\right] z_{ij3}z_{ik3}\right\}. \end{aligned}$$

Temos que

$$\frac{\partial \eta_{i3}}{\partial \beta_{j3}} = z_{ij3} \qquad \text{e} \qquad \frac{\partial \eta_{i3}}{\partial \beta_{k3}} = z_{ik3}.$$

Sob condições de regularidade, $E\left(\partial \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y})/\partial p_{11i}\right) = 0$. Portanto, obtém-se que o componente (j,k) de $J_{\boldsymbol{\beta}_3\boldsymbol{\beta}_3}$ é dado por

$$\begin{aligned} -E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j3}\partial \beta_{k3}}\right) &= \sum_{i=1}^{n}\left\{\left[-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}^2}\right]\left(\frac{\partial p_{11i}}{\partial \eta_{i3}}\right)^2\right] z_{ij3}z_{ik3}\right\} \\ &= \sum_{i=1}^{n} f_{p_{11i}}(l_{p_{11i}})^2 z_{ij3}z_{ik3}, \end{aligned}$$

em que

$$\begin{aligned} fp_{11i} &= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{11i}^2}\right] \\ &= \frac{(d_{31i})^2}{a_{31i}} - \frac{l_{32i}a_{32i}-(d_{32i})^2}{a_{32i}} + 2p_{1i} \\ &\quad - \frac{l_{13i}a_{13i}-(d_{13i})^2}{a_{13i}} - \frac{l_{23i}a_{23i}-(d_{23i})^2}{a_{23i}}, \end{aligned}$$

sendo que l_{ij} representa a derivadas de segunda ordem das probabilidades a_{ij} com relação a p_{11i}, ou seja

$$\begin{aligned} l_{31i} &= 0 \\ l_{32i} &= \left\{p_{0i}\left[2(1-p_{00i})^2 - 4p_{00i}(1-p_{00i})\right] + \right. \\ &\quad \left. p_{1i}\left[4p_{00i}(1-p_{00i}) - 2p_{00i}^2\right]\right\} \\ l_{13i} &= \{p_{0i}[2(1-p_{00i})] + p_{1i}[-4(1-p_{00i})]\} \\ l_{23i} &= \left\{p_{0i}\left[-4(1-p_{00i})^2 + 2p_{00i}(1-p_{00i})\right] + \right. \\ &\quad \left. p_{1i}\left[2(1-p_{00i})^2 - 4p_{00i}(1-p_{00i})\right]\right\}. \end{aligned}$$

Logo, podemos escrever $J_{\boldsymbol{\beta}_3\boldsymbol{\beta}_3}$ na forma matricial conforme apresentado em (4.4). De forma análoga,

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j1}\partial \beta_{k2}} &= \sum_{i=1}^{n} \frac{\partial^2 \ell_i(p_{0i},p_{00i})}{\partial \beta_{j1}\partial \beta_{k2}} \\ &= \sum_{i=1}^{n} \frac{\partial^2 \ell_i(p_{0i},p_{00i})}{\partial p_{0i}\partial p_{00i}} \frac{\partial p_{0i}}{\partial \eta_{i1}} \frac{\partial p_{00i}}{\partial \eta_{i2}} \frac{\partial \eta_{i1}}{\partial \beta_{j1}} \frac{\partial \eta_{i2}}{\partial \beta_{k2}}. \end{aligned}$$

Portanto, obtemos que o termo (j,k) de $J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_2}$ é dado por

$$\begin{aligned} -E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial \beta_{j1}\partial \beta_{k2}}\right) &= -\sum_{i=1}^{n} E\left(\frac{\partial^2 \ell_i(p_{0i},p_{00i})}{\partial p_{0i}\partial p_{00i}}\right) \frac{\partial p_{0i}}{\partial \eta_{i1}} \frac{\partial p_{00i}}{\partial \eta_{i2}} z_{ij1}z_{ik2} \\ &= \sum_{i=1}^{n} f_{i_{p_0p_{00}}} l_{p_{0i}} l_{p_{00i}} z_{ij1}z_{ik2}, \end{aligned}$$

em que

$$\begin{aligned} f_{i_{p_0p_{00}}} &= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}\partial p_{00i}}\right] \\ &= -\frac{\delta_{31i}a_{31i}-b_{31i}c_{31i}}{a_{31i}} - \frac{\delta_{32i}a_{32i}-b_{32i}c_{32i}}{a_{32i}} \\ &\quad - \frac{\delta_{13i}a_{13i}-b_{13i}c_{13i}}{a_{13i}} - \frac{\delta_{23i}a_{23i}-b_{23i}c_{23i}}{a_{23i}}, \end{aligned}$$

sendo que δ_{ij} representa a derivada de b_{ij} com relação a p_{00i}, dadas por

$$\begin{aligned}
\delta_{31i} &= 2(1-2p_{00i})(1-p_{11i}) - 2p_{00i}(1-p_{11i}) \\
\delta_{32i} &= \{2p_{11i}(1-p_{11i})(1-2p_{00i}) + 2(1-p_{11i})^2(p_{00i}-1) \\
&\quad -2p_{00i}p_{11i}(1-p_{11i}) - 2(1-p_{11i})^2(1-2p_{00i})\} \\
\delta_{13i} &= p_{11i}(2-3p_{11i}) \\
\delta_{23i} &= \{[4p_{11i}(1-p_{11i})(p_{00i}-1) + p_{11i}^2(1-2p_{00i})] - \\
&\quad [2p_{11i}(1-p_{11i})(1-2p_{00i}) + 2(1-p_{11i})^2(p_{00i}-1)]\}.
\end{aligned}$$

Logo, podemos escrever $J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_2}$ na forma matricial conforme apresentado em (4.4). De maneira equivalente,

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial\beta_{j1}\partial\beta_{k3}} &= \sum_{i=1}^{n} \frac{\partial^2 \ell_i(p_{0i},p_{11i})}{\partial\beta_{j1}\partial\beta_{k3}} \\
&= \sum_{i=1}^{n} \frac{\partial^2 \ell_i(p_{0i},p_{11i})}{\partial p_{0i}\partial p_{11i}} \frac{\partial p_{0i}}{\partial\eta_{i1}} \frac{\partial p_{11i}}{\partial\eta_{i3}} \frac{\partial\eta_{i1}}{\partial\beta_{j1}} \frac{\partial\eta_{i3}}{\partial\beta_{k3}}.
\end{aligned}$$

Portanto, obtemos que o termo (j,k) de $J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_3}$ é dado por

$$\begin{aligned}
-E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial\beta_{j1}\partial\beta_{k3}}\right) &= -\sum_{i=1}^{n} E\left(\frac{\partial^2 \ell_i(p_{0i},p_{11i})}{\partial p_{0i}\partial p_{11i}}\right) \frac{\partial p_{0i}}{\partial\eta_{i1}} \frac{\partial p_{11i}}{\partial\eta_{i3}} z_{ij1}z_{ik3} \\
&= \sum_{i=1}^{n} f_{i_{p_0p_{11}}} l_{p_{0i}} l_{p_{11i}} z_{ij1}z_{ik3},
\end{aligned}$$

em que

$$\begin{aligned}
f_{i_{p_0p_{11}}} &= -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{0i}\partial p_{11i}}\right] \\
&= -\frac{f_{31i}a_{31i} - b_{31i}d_{31i}}{a_{31i}} - \frac{f_{32i}a_{32i} - b_{32i}d_{32i}}{a_{32i}} \\
&\quad -\frac{f_{13i}a_{13i} - b_{13i}d_{13i}}{a_{13i}} - \frac{f_{23i}a_{23i} - b_{23i}d_{23i}}{a_{23i}},
\end{aligned}$$

sendo que f_{ij} representa a derivada de b_{ij} com relação a p_{11i}, ou seja

$$\begin{aligned}
f_{31i} &= p_{00i}(3p_{00i}-2) \\
f_{32i} &= \{2p_{00i}(1-p_{00i})(1-2p_{11i}) + 2(1-p_{00i})^2(p_{11i}-1) \\
&\quad -p_{00i}^2(1-2p_{11i}) - 4p_{00i}(1-p_{00i})(p_{11i}-1)\} \\
f_{13i} &= \{2p_{11i}(1-p_{00i}) - 2(1-p_{00i})(1-2p_{11i})\} \\
f_{23i} &= \{[2(1-p_{00i})^2(1-2p_{11i}) + 2p_{00i}p_{11i}(1-p_{00i})] - \\
&\quad [2p_{00i}(1-p_{00i})(1-2p_{11i}) + 2(1-p_{00i})^2(p_{11i}-1)]\}.
\end{aligned}$$

Logo, podemos escrever $J_{\boldsymbol{\beta}_1\boldsymbol{\beta}_3}$ na forma matricial conforme apresentado em (4.4).

De maneira semelhante,

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial\beta_{j2}\partial\beta_{k3}} &= \sum_{i=1}^{n}\frac{\partial^2\ell_i(p_{00i},p_{11i})}{\partial\beta_{j2}\partial\beta_{k3}}\\ &= \sum_{i=1}^{n}\frac{\partial^2\ell_i(p_{00i},p_{11i})}{\partial p_{00i}\partial p_{11i}}\frac{\partial p_{00i}}{\partial\eta_{i2}}\frac{\partial p_{11i}}{\partial\eta_{i3}}\frac{\partial\eta_{i2}}{\partial\beta_{j2}}\frac{\partial\eta_{i3}}{\partial\beta_{k3}}.\end{aligned}$$

Portanto, obtemos que o termo (j,k) de $J_{\boldsymbol{\beta}_2\boldsymbol{\beta}_3}$ é dado por

$$\begin{aligned}-E\left(\frac{\partial^2\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial\beta_{j2}\partial\beta_{k3}}\right) &= -\sum_{i=1}^{n}E\left(\frac{\partial^2\ell_i(p_{00i},p_{11i})}{\partial p_{00i}\partial p_{11i}}\right)\frac{\partial p_{00i}}{\partial\eta_{i2}}\frac{\partial p_{11i}}{\partial\eta_{i3}}z_{ij2}z_{ik3}\\ &= \sum_{i=1}^{n}f_{i_{p_{00}p_{11}}}l_{p_{00i}}l_{p_{11i}}z_{ij2}z_{ik3},\end{aligned}$$

em que

$$\begin{aligned}f_{i_{p_{00}p_{11}}} &= E\left[\frac{\partial^2\ell(\boldsymbol{\theta}|(\boldsymbol{x},\boldsymbol{y}))}{\partial p_{00i}\partial p_{11i}}\right]\\ &= -\frac{q_{31i}a_{31i}-c_{31i}d_{31i}}{a_{31i}}-\frac{q_{32i}a_{32i}-c_{32i}d_{32i}}{a_{32i}}\\ &\quad -\frac{q_{13i}a_{13i}-c_{13i}d_{13i}}{a_{13i}}-\frac{q_{23i}a_{23i}-c_{23i}d_{23i}}{a_{23i}},\end{aligned}$$

sendo que q_{ij} representa a derivada c_{ij} com relação a p_{11i}, ou seja

$$\begin{aligned}q_{31i} &= p_{0i}[-2(1-2p_{00i})]+p_{1i}[-2p_{00i}]\\ q_{32i} &= \{p_{0i}[2(1-2p_{11i})(1-2p_{00i})+4(p_{11i}-1)(p_{00i}-1)]\\ &\quad +p_{1i}[4(p_{11i}-1)(1-2p_{00i})+2p_{00i}(1-2p_{11i})]\}\\ q_{13i} &= p_{0i}\{-2p_{11i}\}+p_{1i}\{-2(1-2p_{11i})\}\\ q_{23i} &= \{p_{0i}[4(1-2p_{11i})(p_{00i}-1)+2p_{11i}(1-2p_{00i})]+\\ &\quad p_{1i}[2(1-2p_{11i})(1-2p_{00i})+4(p_{11i}-1)(p_{00i}-1)]\}.\end{aligned}$$

Logo, podemos escrever $J_{\boldsymbol{\beta}_2\boldsymbol{\beta}_3}$ na forma matricial conforme apresentado em (4.4).


ICMC
USP
SÃO CARLOS

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PROGRAMA INTERINSTITUCIONAL DE PÓS GRADUAÇÃO
EM ESTATÍSTICA

Thiago Souza de Melo

MODELOS DE REGRESSÃO PARA PLACARES DE JOGOS DIVIDIDOS EM *SETS*

São Carlos - SP

2026