Felipe Nelson Victoria

Um modelo de previsão de resultados de partidas de futebol: Flamengo deveria ser o campeão em 2009?

Niterói - RJ, Brasil

21 de dezembro de 2022

Universidade Federal Fluminense

Felipe Nelson Victoria

Um modelo de previsão de resultados de partidas de futebol: Flamengo deveria ser o campeão em 2009?

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Profa. Dra. Karina Yuriko Yaginuma

Niterói - RJ, Brasil

21 de dezembro de 2022

Universidade Federal Fluminense

Felipe Nelson Victoria

Um modelo de previsão de resultados de partidas de futebol: Flamengo deveria ser o campeão em 2009?

Monografia de Projeto Final de Graduação sob o título *"Um modelo de previsão de resultados de partidas de futebol: Flamengo deveria ser o campeão em 2009?"*, defendida por Felipe Nelson Victoria e aprovada em 21 de dezembro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Folha de assinaturas

Profa. Dra. Karina Yuriko Yaginuma
Departamento de Estatística – UFF

Prof. Dr. Hugo Henrique Kegler dos Santos
Departamento de Estatística – UFF

Prof. Dr. Marco Aurélio dos Santos Sanfins
Departamento de Estatística – UFF

Niterói, 21 de dezembro de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

V642m Victoria, Felipe Nelson
Um modelo de previsão de resultados de partidas de futebol: Flamengo deveria ser o campeão em 2009? / Felipe Nelson Victoria. - 2022.
31 f.

Orientador: Karina Yuriko Yaginuma.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Método Soma e Diferença 0. 2. Modelo de predição. 3. Futebol. 4. Produção intelectual. I. Yuriko Yaginuma, Karina, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O trabalho tem como principal objetivo desenvolver um modelo capaz de predizer os resultados para as partidas de futebol e o propósito secundário de demonstrar se o Flamengo era o time favorito para ser campeão brasileiro no ano de 2009. Para a realização do trabalho foi utilizado o método da Soma e Diferença 0, onde é assumido que X e Y são os números de gols marcados em uma dada partida pelos times mandante e visitante, respectivamente. Mais ainda, é admitido que X e Y são variáveis aleatórias independentes e ambas seguem uma distribuição de Poisson com parâmetros λ_X e λ_Y. Os ajustes dos modelos são feitos por meio das informações de quais são os times mandantes e visitantes e do número de gols marcados por cada um deles nas partidas do primeiro turno de cada campeonato analisado. Em posse do parâmetros, são feitas simulações para o segundo turno do torneio. Dessa forma torna-se viável apontar os times que mais vezes ganharam os campeonatos e assim, afirmar se um time é ou não um dos favoritos para se consagrar como campeão do Campeonato Brasileiro. O modelo desenvolvido foi capaz de prever os resultados finais das partidas disputadas na segunda metade dos campeonatos, apontando se houve uma equipe vencedora ou se o jogo terminou empatado e indicou que o Flamengo não era uma das equipes favoritas ao título em 2009.

Palavras-chave: Método Soma e Diferença 0. Modelo de predição. Futebol.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O futebol é um dos esportes mais populares ao redor de todo o mundo e é possível encontrá-lo liderando diversas listas que enumeram os esportes mais praticados pelas pessoas. Ele já é um esporte consolidado em diversos paises da Europa e da América do Sul, porém sua prática vem ganhando força em locais como China e Estados Unidos. Segundo a FIFA, 3,5 bilhões de pessoas assistiram pelo menos um jogo do mundial de 2018. Somado a sua popularidade, tem-se o ardor de seus espectadores por seus times e pelo jogo em si, tornando-o um dos esportes que mais movimenta dinheiro em todo o planeta. Entretanto, antes de explorar o mercado bilionário que esse jogo movimenta, é importante que o leitor saiba um pouco de sua história e de como o esporte se tornou o que é hoje.

Os historiadores não conseguiram encontrar com exatidão uma data para o surgimento do esporte, todavia, existem diversas referências durante toda a Antiguidade de atividades "parecidas" com o futebol. Na China, por volta de 2.600 a.C, as tribos chinesas praticavam o ritual denominado "TsuTsu" onde as cabeças dos chefes inimigos perdedores eram chutadas pelos líderes ganhadores. Já no Japão, há o surgimento do "Kemari", uma atividade que consistia em controlar a bola com os pés. Na Europa, mais precisamente na Itália, apareceu a primeira versão com uma forma mais próxima ao que existe atualmente. Era chamado de "*Calcio Storico*" e consistia em um jogo onde equipes compostas por 30 integrantes cada, deveriam tentar atravessar a bola pela trave adversária.

De acordo com (VALERIANO, 2020) o esporte foi levado para Inglaterra e lá, em 1863, a reunião "*Freemason's Tavern*" fundou a "*Football Association*" (FA), instituição que elaborou as primeiras regras para o jogo, campeonatos e partidas oficiais e, graças a esse fato, foi possível que ele fosse difundido em outros países. No Brasil, o futebol foi trazido por Charles Miller, em 1894.

Com a popularização do futebol ao redor de todo mundo, não é de se espantar que junto a esse fato tenha crescido também o mercado de apostas em torno de suas partidas.

As apostas em esportes são datadas muito antes mesmo da profissionalização e propagação do futebol. Existem relatos de que os gregos acreditavam que o jogo era uma condição imprescindível para o alcance elevado do estado de espírito, que os romanos faziam apostas em jogos de gladiadores e durante a Revolução Industrial eram comum as apostas em corridas de cavalos.

Em seu artigo (KOKENY, 2021) expõe que em meados do ano 1995 começou a se desenvolver o modelo de apostas "*online*", onde empresas criaram os sites que possibilitavam a elaboração de apostas utilizando a internet. No início dos anos 2000, surgiram as primeiras apostas ao vivo, onde é possível fazer apostas enquanto um determinado jogo está acontecendo.

Hoje em dia, o mercado de apostas só tem crescido cada vez mais no mundo. Foi realizada uma pesquisa pela *Grand View Research* e, em 2020, esse mercado foi avaliado em US$ 59,6 bilhões e para 2027 é esperado que o valor chegue a US$ 127,3 bilhões.

O Brasil vem seguindo os mesmo passos do restante dos grandes mercados e, com a nova regulamentação sancionada pelo ex-presidente Michel Temer, os números do mercado de aposta brasileiro só crescem. Segundo o *Mktesportivo*, o faturamento desse mercado, em 2020, foi de R$ 7 bilhões.

A motivação para a elaboração deste trabalho foi a soma de alguns fatores. O primeiro é a paixão em si pelo futebol, o segundo fator foi o polêmico campeonato brasileiro de 2009, onde muitas pessoas afirmam que o Flamengo somente foi campeão por ter sido ajudado por diversos times que não estavam na briga pelo campeonato, e, por último, o potencial mercado de apostas que o Brasil tem se mostrado. Dessa forma, procurou-se algum tipo de modelo que fosse capaz de prever os resultados das partidas com uma taxa de acerto razoável e, por conseguinte, tentar determinar se a equipe do Flamengo merecia ser campeã em 2009.

Durante as pesquisa preliminares, foi possível perceber que existe uma vasta quantidade de estudos para essa área de previsão de resultados de partidas. Para essa monografia, foi utilizado como referência o modelo proposto por (SUZUKI; TAVARES, 2015), onde os autores fazem uso do método da Soma e Diferença 0 (SD0) para elaborar um modelo de previsão de resultados.

Diversos modelos para a previsão dos resultados das partidas de futebol assumem que os gols marcados pelos times em uma dada partida são duas variáveis aleatórias independentes e ambas seguem uma distribuição de Poisson. Partindo disso, no trabalho são criadas outras duas variáveis, a soma e a diferença de gols dessa partida e define-se um modelo de regressão linear para os parâmetros relacionados ao poder defensivo e ofensivo de cada time.

1.1 Objetivos

Foram pensados em dois objetivos para o presente projeto. O primeiro é o de desenvolver um modelo que seja capaz de fazer previsões sobre o número de gols marcados em uma partida e, dessa forma, conseguir simular partidas para predizer em qual colocação a equipe irá terminar o campeonato. O segundo objetivo é o de tentar mostrar que o time do Flamengo tinha, de fato, as maiores chances de ser a campeã do Campeonato Brasileiro de 2009.

1.2 Organização

A apresentação da monografia está organizada em 3 partes, sendo o Capítulo 2 é uma introdução sobre o tema, mostrando conceitos importantes e aborda como foi a contrução do modelo. No Capítulo 3, são apresentados os resultados obtidos por meio do modelo elaborado e no Capítulo 4 é feito uma breve conclusão sobre os objetivos propostos no trabalho.

2 Materiais e Métodos

O capítuo apresenta O capítulo apresenta a base de dados que será utilizada no trabalho, conceitos importantes e a definição do modelo para a previsão de resultados de partidas de futebol.

2.1 Dados

A base de dados utilizada no trabalho foi coletada manualmente no site da RSSSF Brasil - Resultados Históricos ⟨https://rsssfbrasil.com/historical.htm⟩ e nela estão contidos os resultados de todas as partidas dos campeonatos brasileiros compreendidos entre os anos de 2005 e 2009.

É importante ressaltar que todos os campeonatos utlizados no estudo foram disputados no modelo de pontos corridos. Isto quer dizer que, se em uma partida houver um time vencerdor para o confronto, ao ganhador serão atribuidos 3 pontos e ao perdedor não terá pontuação acrescida. Em caso de empate, ambas as equipes irão receber 1 ponto cada. Ao fim do campeonato, o time que possuir maior quantidade de pontos será declarado campeão.

Ela é composta por 1.982 observações e 9 variáveis, dividas em 5 tabelas referentes a cada temporada, onde cada campeonato é compostos por 380 partidas e possui 20 times participantes, a exceção está no ano de 2005, que conta com 462 partidas e 22 times.

Com o intuito de facilitar a compreensão sobre a base de dados anteriomente mencionada, foi criada a tabela 1. Nela são expostas as variáveis e suas respectivas representações, em que as informações contidas são referentes a i-ésima partida do campeonato. Além dela é também apresentada a Figura 1 na qual é mostrado um exemplo de como são os dados dentro do programa RStudio.

Tabela 1: Variáveis da base de dados.

Variável	Representação
Partida	Representa o número partida
Time Mandante	Representa o time mandante
Gols Mandante	Representa o número de gols marcados pelo time mandante
Gols Visitante	Representa o número de gols marcados pelo time visitante
Time Visitante	Representa o time visitante
Vencedor	Variável que assume os valores ”mandante”, ”visitante”e ”empate”
Time Ganhador	Representa o time ganhador
Time Perdedor	Representa o time perdedor
Total de Gols	Determina o total de gols

Figura 1: Exemplo da base de dados

	Partida	Mandante	GolsMandante	GolsVisitante	Visitante	Vencedor	Time Ganhador	Time Perdedor	Total Gols
1	1	Sport	1	1	Barueri	Empate	Empate	Empate	2
2	2	Palmeiras	2	1	Coritiba	Mandante	Palmeiras	Coritiba	3
3	3	Avai	2	2	Atletico_MG	Empate	Empate	Empate	4
4	4	Corinthians	0	1	Internacional	Visitante	Internacional	Corinthians	1
5	5	Fluminense	1	0	Sao_Paulo	Mandante	Fluminense	Sao_Paulo	1
6	6	Cruzeiro	2	0	Flamengo	Mandante	Cruzeiro	Flamengo	2
7	7	Atletico_PR	0	2	Vitoria	Visitante	Vitoria	Atletico_PR	2
8	8	Gremio	1	1	Santos	Empate	Empate	Empate	2
9	9	Santo_Andre	1	1	Botafogo	Empate	Empate	Empate	2
10	10	Goias	3	3	Nautico	Empate	Empate	Empate	6

2.2 Metodologia

Tendo como objeto de estudo a série A dos campeonatos brasileiros compreendidos entre os anos de 2005 a 2009, o trabalho pretende avaliar a existência de relação entre os gols feitos em uma determinada partida de futebol com as variáveis: poder ofensivo e poder defensivo, a fim de prever qual será o time vencedor do campeonato e avaliar se a equipe do Flamengo merecia ou não ser a campeã no ano de 2009, isto é, se após a realização das simulações ele seria a equipe com maior probabilidade de ser campeão ou não. Dessa forma, a monografia é desenvolvida baseada no modelo proposto por (SUZUKI; TAVARES, 2015) onde é aplicado método Soma e Diferença 0 (SD0) conjuntamente com o uso dos modelos lineares, para a estimação das probabilidades de ganho, perda ou empate de um time em uma determinada partida.

O método SD0 pode ser resumido da seguinte maneira. Seja X o número de gols marcados pelo time mandante e Y número de gols marcados pelo time visitante para uma determinada partida. Supondo que X e Y são duas variáveis aleatórias independentes e ambas com distribuição de Poisson, então suas as médias λ_X e λ_Y representam os números de gols esperados para cada time naquela partida em questão. A estimação dos parâmetros λ_X e λ_Y é feita por meio da esperança da soma $X+Y$ e da diferença $X-Y$, respectivamente. Determinados esses valores é possível determinar as chances de vitória de cada time em uma partida.

Antes de iniciar a explicação da metodologia utilizada no trabalho, será realizada uma breve explanação de alguns conceitos importantes para o bom entendimentos no modelo proposto.

2.2.1 A distribuição de Poisson Bivariada

Um conceito amplamente difundido em diversos artigos e estudos é a suposição de que o número de gols marcados por uma determinada equipe de futebol em uma partida possua uma distribuição de Poisson. Dessa forma, é possível definir uma distribuição discreta bivariada capaz de modelar a quantidade de gols marcados por duas equipes em um jogo. Então, tem-se um vetor aleatório (X, Y) com suporte $\mathbb{N}^2$, onde X e Y são os números de gols marcados pelos times mandantes e visitantes, respectivamente. Além disso, é possível afirmar que elas são variáveis aleatórias independentes, visto que o total gols marcados por uma equipe não é influenciado pelo do outro.

Da mesma forma que (ARRUDA, 2000) utiliza em sua tese a distribuição de Poisson Bivariada "de Holgate", o modelo também fará uso dela. Essa distribuição é construída por meio de três processos de Poisson independentes, sendo Z_X, Z_{XY} e Z_Y os números de ocorrências de cada processo durante um período de duração comum e essa variáveis aleatórias possuem médias λ_X, λ_{XY} e λ_Y, respectivamente. Dessa forma, as variáveis $X = Z_X + Z_{XY}$ e $Y = Z_Y + Z_{XY}$ têm distribuição conjunta de Poisson bivariada. É perceptível que X e Y têm uma distribuição de Poisson com médias $\lambda_X + \lambda_{XY}$ e $\lambda_Y + \lambda_{XY}$.

A função de probabilidade conjunta de X e Y é dada por:

$$P(X = x, Y = y) = e^{-(\lambda_X + \lambda_Y + \lambda_{XY})} \times \sum_{i=0}^{min(X,Y)} \frac{\lambda_X^{x-i} \lambda_Y^{y-i} \lambda_{XY}^{i}}{(x-i)!(y-i)!}.$$

E as marginais são dadas por:

$$P(X = x) = \sum_{y=0}^{\infty} P(X = x, Y = y) = \frac{e^{\lambda_X} \times (\lambda_X)^x}{x!}, \text{ para algum } \lambda_X > 0,$$

$$P(Y = y) = \sum_{x=0}^{\infty} P(X = x, Y = y) = \frac{e^{\lambda_Y} \times (\lambda_Y)^y}{y!}, \text{ para algum } \lambda_Y > 0.$$

2.2.2 Matriz Inversa de Moore-Penrose

Em sua tese (ALVES, 1990) explica que o conceito de uma inversa generalizada de uma matriz $A_{m\times n}$ foi elaborado por Moore (1920) e desenvolvido por Penrose (1955). Essa matriz inversa, denotada por $A^{\ominus}_{n\times m}$, se faz necessária quando a matriz $A_{m\times n}$ é posto coluna completo e o sistema de equações normais é indeterminado. Em outras palavras, $A^{\ominus}_{n\times m}$ é utilizada para calcular uma solução de "melhor ajuste"(por mínimos quadrados) para um sistema de equações lineares que não possui uma solução ou para encontrar a solução da norma mínima (euclidiana) para um sistema de equações lineares com múltiplas soluções.

Em seu artigo (SAMEJIMA, 2020) explica que a inversa de Moore-Penreose é uma inversa especial, já que é viável garantir a unicidade dela. Ele apresenta as seguintes definições:

Definição 2.1 *A matriz $A^{\ominus}$ é inversa generalizada de Moore-Penrose de A se:*

- $A^{\ominus}A = I$;
- $A^{\ominus} = (A^TA)^{-1}A^T$;
- *Ela é uma inversa generalizada de A, isto é,* $AA^{\ominus}A = A$;
- *Ela é reflexiva, isto é,* $A^{\ominus}AA^{\ominus} = A^{\ominus}$;
- $A^{\ominus}A$ *e* $AA^{\ominus}$ *são matrizes simétricas.*

Dado um modelo linear

$$\mathbf{Y} = X\beta + \epsilon,$$

com X matriz de posto incompleto. Logo,

$$\hat{\mathbf{b}} = GX^T\mathbf{y}$$

será um estimador de mínimos quadrados para β.

Uma outra maneira de visualizar essa ideia é supor que a matriz A não é uma matriz quadrada, dessa forma, A^{-1} não pode ser calculada, então se faz uso da inversa de Moore-Penrose $A^{\ominus}$ para encontrar as soluções aproximadas para o sistema.

Seja $\vec{y} = A\vec{x}$ um sistema de equações lineares em que o vetor $\vec{y}$ e a matriz A são conhecidos e o número de linhas m é maior do que o número de colunas n. Do sistema acima, tem-se que:

$$\begin{aligned} a_{1,1}x_1 + a_{2,1}x_1 + ... + a_{1,n}x_n = y_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + ... + a_{2,n}x_n = y_2 \\ ... \\ a_{m,1}x_1 + a_{m,2}x_1 + ... + a_{m,n}x_n = y_n \end{aligned} .$$

Dessa forma, o sistema pode ser reescrito da forma

$$\vec{y} = A\vec{x}$$

$$\Rightarrow A^{\ominus}\vec{y} \approx A^{\ominus}A\vec{x}$$

$$\Rightarrow A^{\ominus}\vec{y} \approx I\vec{x}$$

$$\Rightarrow A^{\ominus}\vec{y} \approx \vec{x}$$

É importante notar que a solução para o sistema será um resultado aproximado, já que não é possível encontrar um valor exato.

2.2.3 Famílias de Métodos Soma e Diferença (SD)

Fazendo uso das propriedades da distribuição Poisson Bivariada "de Holgate", percebe-se que a esperança da soma e da diferença entre X e Y são dadas por:

$$E[X-Y] = E[X] - E[Y] = (\lambda_X + \lambda_{XY}) - (\lambda_Y + \lambda_{XY}) = \lambda_X - \lambda_Y,$$

$$E[X+Y] = E[X] + E[Y] = (\lambda_X + \lambda_{XY}) + (\lambda_Y + \lambda_{XY}) = \lambda_X + \lambda_Y + 2\lambda_{XY}.$$

Esse dois resultados são muito importantes paras as famílias Soma e Diferença e a distinção entre essas famílias se dá apenas pela estimação do parâmetro λ_{XY}, que pode ser entendido como a covariância entre as variáveis X e Y.

A seguir é apresentado um caso particular da família de métodos Soma e Diferença que será utilizado no trabalho.

2.2.3.1 Método Soma e Diferença 0 (SD0)

O Método Soma e Diferença 0 é uma das alternativas mais simplificadas da família de métodos soma e diferença, pois nesse método é assumido que o valor da covariância (λ_{XY}) é igual a zero. Esse resultado só é possível porque é admitida a independência entre o número de gols marcados pelos times mandante e visitantes.

Os parâmetros λ_X e λ_Y podem ser obtidos das expressões apresentadas na seção 2.2.2 das Famílias de Métodos Soma e Diferença. Ou seja, através do seguinte sistema de equações:

$$\begin{cases} E[X-Y] = \lambda_X - \lambda_Y \\ E[X+Y] = \lambda_X + \lambda_Y \end{cases}.$$

a solução do sistema é dada por:

$$\begin{cases} \lambda_X = \frac{E[X+Y]+E[X-Y]}{2} \\ \lambda_Y = \frac{E[X+Y]-E[X-Y]}{2} \end{cases}. \tag{2.1}$$

Como foi elucidado em 2.2.1 tem-se que X e Y são duas variáveis aleatórias independentes e ambas seguem uma distribuição de Poisson com parâmetros λ_X e λ_Y, respectivamente. E, mais ainda, X representa o número de gols marcados pelo time mandante e Y representa o número de gols marcados pelo time visitante em uma determinada partida.

Note que os valores de λ_X e λ_Y representam os resultados esperados para o número de gols marcados em uma determinada partida pelos times mandante e visitante. Mais do que isso, eles podem ser expressos como funções das médias das variáveis $X+Y$ e $X-Y$.

Assim sendo, definem-se duas variáveis $U = X+Y$ e $V = X-Y$, em que U representa a soma e V a diferença dos gols para uma partida qualquer. Desse conceito, a estimação dos parâmetros λ_X e λ_Y pode ser feita através da obtenção dos valores estimados das médias $\mu = E[U]$ e $\nu = E[V]$. Esses valores estimados estão relacionados a fatores de poder ofensivo do time mandante e o poder defensivo do time visitante.

Logo, o sistema de equações visto em 2.1 pode ser reescrito da seguinte forma:

$$\hat{\lambda}_X = \frac{\hat{E}[X+Y]+\hat{E}[X-Y]}{2} = \frac{\hat{E}[U]+\hat{E}[V]}{2} = \frac{\hat{\mu}+\hat{\nu}}{2}; \tag{2.2}$$

$$\hat{\lambda}_Y = \frac{\hat{E}[X+Y]-\hat{E}[X-Y]}{2} = \frac{\hat{E}[U]-\hat{E}[V]}{2} = \frac{\hat{\mu}-\hat{\nu}}{2}. \tag{2.3}$$

Então, para encontrar os valores dos estimadores $\hat{\mu}$ e $\hat{\nu}$ são definidos dois modelos lineares.

$$U_i = \mu_i + \epsilon_i \text{ , com } \mu_i = \beta_{h[i]} + \psi_{\alpha[i]} \tag{2.4}$$

$$V_i = \nu_i + \epsilon_i^{'} \text{ , com } \nu_i = \beta_{h[i]}^{'} + \psi_{\alpha[i]}^{'} \tag{2.5}$$

onde, para cada partida, i = 1, 2, 3, ...n:

- $\beta_{h[i]}$ e $\beta_{h[i]}^{'}$ referem-se ao poder ofensivo do time $h[i]$;

- $\psi_{\alpha[i]}$ e $\psi'_{\alpha[i]}$ referem-se ao poder defensivo do time $\alpha[i]$;
- ϵ_i e ϵ'_i são os erros aleatórios independentes com média zero.

É importante notar que, diferentemente do modelo apresentado em (SUZUKI; TAVARES, 2015), o modelo utilizado considera que o mandante da partida sempre está jogando em casa, mesmo quando o jogo é realizado em outro estádio. Isso se deve ao fato de que, em alguns casos, os times precisam ou optam por jogarem suas partidas em um campo diferente.

Assim sendo, optou-se por usar duas matrizes indicadoras que informam quais são os times mandantes e visitantes da partida. Portanto, as matrizes denominadas $\mathbf{B}_{n\times m}$ e $\mathbf{P}_{n\times m}$ indicam o time mandante e visitante, respectivamente. Veja que n é a quantidade de partidas realizadas durante o 1º turno do campeonato observado e m é o número de times partipantes.

Logo, as equações 2.4 e 2.5 podem ser reescritas da seguinte forma:

$$\underline{U} = \mathbf{B}_{n\times m} \times \underline{\beta} + \mathbf{P}_{n\times m} \times \underline{\psi} + \underline{\epsilon}; \tag{2.6}$$

$$\underline{V} = \mathbf{B}_{n\times m} \times \underline{\beta'} + \mathbf{P}_{n\times m} \times \underline{\psi'} + \underline{\epsilon'}. \tag{2.7}$$

Onde:

- $\underline{U} = (U_1, U_2, U_3, ..., U_n)$ e $\underline{V} = (V_1, V_2, V_3, ..., V_n)$
- $\underline{\beta} = (\beta_1, \beta_2, \beta_3, ..., \beta_T)$ e $\underline{\beta'} = (\beta'_1, \beta'_2, \beta'_3, ..., \beta'_T)$ referem-se ao coeficientes de poder ofensivo do time mandante, com $T \in (1, 2, 3, ..., m)$;
- $\mathbf{B}_{n\times m} = \begin{cases} 1, \text{ caso o time seja o mandante} \\ 0, \text{ caso contrário} \end{cases};$
- $\underline{\psi} = (\psi_1, \psi_2, \psi_3, ..., \psi_T)$ e $\underline{\psi'} = (\psi'_1, \psi'_2, \psi'_3, ..., \psi'_T)$ referem-se ao coeficientes de poder defensivo do time visitante, com $T \in (1, 2, 3, ..., m)$;
- $\mathbf{P}_{n\times m} = \begin{cases} 1, \text{ caso o time seja o visitante} \\ 0, \text{ caso contrário} \end{cases};$
- $\underline{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, , ..., \epsilon_n)$ e $\underline{\epsilon'} = (\epsilon'_1, \epsilon'_2, \epsilon'_3, , ...\epsilon'_n)$ são os erros aleatórios independentes com média zero.

Um detalhe importante que deve ser destacados para a definição dos modelos é o fato dos métodos SD se basearem em modelos lineares sem interceptos, isso se já que todas as partidas sempre começam com o placar de 0x0.

Adiante é demonstrado um esquema das partidas.

Partida	*Mandante*	*Visitante*	U_i	V_i
1	β_1	ψ_1	U_1	V_1
2	β_2	ψ_2	U_2	V_2
3	β_3	ψ_3	U_3	V_3
...	...	...	...	...
$\frac{n}{2}$	$\beta_{\frac{n}{2}}$	$\psi_{\frac{n}{2}}$	$U_{\frac{n}{2}}$	$V_{\frac{n}{2}}$
$\frac{n}{2}+1$	$\beta'_{\frac{n}{2}+1}$	$\psi'_{\frac{n}{2}+1}$	$U_{\frac{n}{2}+1}$	$V_{\frac{n}{2}+1}$
$\frac{n}{2}+2$	$\beta'_{\frac{n}{2}+2}$	$\psi'_{\frac{n}{2}+2}$	$U_{\frac{n}{2}+2}$	$V_{\frac{n}{2}+2}$
...	...	...	...	...
n	β'_n	ψ'_n	U_n	V_n

Definidos os dois modelos lineares para U e V, agora é possível encontrar os valores dos coeficientes $\vec{\beta}$, $\vec{\beta}'$, $\vec{\psi}$ e $\vec{\psi}'$. Para isso, é criada uma matriz $\mathbf{A}_{n\times 2m}$ que é a junção entre as matrizes $\mathbf{B}_{n\times m}$ e $\mathbf{P}_{n\times m}$ e os vetores coluna $\underline{\alpha} = (\underline{\beta}, \underline{\psi}) = (\beta_1, \beta_2, \beta_3, ...\beta_T, \psi_1, \psi_2, \psi_3, ..., \psi_T)$ e $\underline{\eta} = (\underline{\beta}', \underline{\psi}') = (\beta'_1, \beta'_2, \beta'_3, ...\beta'_T, \psi'_1, \psi'_2, \psi'_3, ..., \psi'_T)$.

Portanto, os modelos 2.6 e 2.7 podem ser reformulados da seguinte maneira:

$$\underline{U} = \mathbf{A} \times \underline{\alpha} + \underline{\epsilon}; \tag{2.8}$$

$$\underline{V} = \mathbf{A} \times \underline{\eta} + \underline{\epsilon}'. \tag{2.9}$$

Em seu livro , (MONTGOMERY; PECK; VINING, 2013) mostram que tal estimação pode ser feita através dos estimadores de mínimos quadrados e, mais do que isso, eles são obtidos pelas seguintes equações:

$$\hat{\underline{\alpha}} = (A^tA)^{-1}(A^t)\underline{U}; \tag{2.10}$$

$$\hat{\underline{\eta}} = (A^tA)^{-1}(A^t)\underline{V}. \tag{2.11}$$

Encontrados os coeficientes da regressão, o próximo passo é estabelecer os valores dos estimadores $\hat{\mu}$ e $\hat{\nu}$ para então determinar os parâmetros λ_X e λ_X. Eles são definidos pelas equações 2.2 e 2.3.

Relembrando que $X \sim Poisson(\lambda_X = \frac{\hat{\mu}+\hat{\nu}}{2})$ e $Y \sim Poisson(\lambda_Y = \frac{\hat{\mu}-\hat{\nu}}{2})$, então agora é viável de determinar o número de gols marcados pelos times em uma dada partida i.

Para isso, basta apenas utilizar as seguintes probabilidades para encontrar os valores do placar:

$$P(X = x) = \frac{(\hat{\lambda_X}^{x})}{x!} \exp^{-\hat{\lambda_X}}; \tag{2.12}$$

$$P(Y = y) = \frac{(\hat{\lambda_Y}^{y})}{y!} \exp^{-\hat{\lambda_Y}}. \tag{2.13}$$

2.2.4 Exemplo

Para propiciar uma maior clareza sobre como é o funcionamento do método, foi elaborado um exemplo de campeonato fictício, em que os 4 maiores times do Rio de Janeiro se enfrentam e os dois times com as maiores pontuações se enfrentam em uma final de jogo único. Ele é composto de 12 partidas, dividias em 2 turnos e os times são: Botafogo, Flamengo, Fluminense e Vasco.

Abaixo são mostrada as tabelas 2 e 3. A primeira apresenta os resultados de todas as partidas e a segunda a tabela de classificação após os 12 jogos do campeonato.

Tabela 2: Partidas campeonato fictício.

Partida	Mandante	X	Y	Visitante
1	flamengo	1	0	botafogo
2	fluminense	1	1	vasco
3	botafogo	3	1	vasco
4	fluminense	1	4	flamengo
5	flamengo	4	0	vasco
6	botafogo	0	0	fluminense
7	vasco	1	0	botafogo
8	flamengo	2	0	fluminense
9	vasco	1	0	fluminense
10	botafogo	0	1	flamengo
11	fluminense	1	2	botafogo
12	vasco	0	1	flamengo

Tabela 3: Classificação do campeonato fictício.

Posição	Equipe	Pontos	Vitórias	Empates	Derrotas	GP	GC	SG
1	Flamengo	18	6	0	0	13	1	12
2	Botafogo	7	2	1	3	5	5	0
3	Vasco	7	2	1	3	4	9	-5
4	Fluminense	2	0	2	4	3	10	-7

Portanto, a final do campeonato fictício será o confronto entre as equipes do Flamengo e do Botafogo.

Considerando todas as partidas, das equações 2.6 e 2.7, tem-se que:

$$\underline{U} = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 4 \\ 0 \\ 1 \\ 2 \\ 1 \\ 1 \\ 3 \\ 1 \end{bmatrix}, \underline{V} = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -3 \\ 4 \\ 0 \\ 1 \\ 2 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \mathbf{B}_{12\times 4} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{P}_{12\times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\underline{\beta} = \begin{bmatrix} \beta_{bot} \\ \beta_{fla} \\ \beta_{flu} \\ \beta_{vas} \end{bmatrix}, \underline{\psi} = \begin{bmatrix} \psi_{bot} \\ \psi_{fla} \\ \psi_{flu} \\ \psi_{vas} \end{bmatrix}, \underline{\beta'} = \begin{bmatrix} \beta'_{bot} \\ \beta'_{fla} \\ \beta'_{flu} \\ \beta'_{vas} \end{bmatrix}, \underline{\psi'} = \begin{bmatrix} \psi'_{bot} \\ \psi'_{fla} \\ \psi'_{flu} \\ \psi'_{vas} \end{bmatrix},$$

$$\underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} \text{ e } \underline{\epsilon'} = \begin{bmatrix} \epsilon'_1 \\ \epsilon'_2 \\ \epsilon'_3 \\ \epsilon'_4 \\ \epsilon'_5 \\ \epsilon'_6 \end{bmatrix}$$

Para realizar os ajustes dos modelos de regressão linear para a soma e diferença de gols são usadas as equações 2.8 e 2.9. Onde $\mathbf{A}_{12x8}$ é a concatenação das matrizes $\mathbf{B}_{12x4}$ e $\mathbf{P}_{12x4}$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\underline{\alpha} = \begin{bmatrix} \beta_{bot} \\ \beta_{fla} \\ \beta_{flu} \\ \beta_{vas} \\ \psi_{bot} \\ \psi_{fla} \\ \psi_{flu} \\ \psi_{vas} \end{bmatrix} = \begin{bmatrix} 0,42 \\ 1,42 \\ 2,04 \\ 0,29 \\ 0,08 \\ 1,58 \\ -1,54 \\ 0,71 \end{bmatrix} \text{ e } \underline{\eta} = \begin{bmatrix} \beta'_{bot} \\ \beta'_{fla} \\ \beta'_{flu} \\ \beta'_{vas} \\ \psi'_{bot} \\ \psi'_{fla} \\ \psi'_{flu} \\ \psi'_{vas} \end{bmatrix} = \begin{bmatrix} 0,42 \\ 1,42 \\ 0,30 \\ 2,04 \\ 0,08 \\ -1,42 \\ 0,21 \\ 1,96 \end{bmatrix}$$

Os parâmetros são dados pelo vetor $\vec{\lambda} = \lambda_{X_{bot}}, \lambda_{X_{fla}}, \lambda_{X_{flu}}, \lambda_{X_{vas}}, \lambda_{Y_{bot}}, \lambda_{Y_{fla}}, \lambda_{Y_{flu}}, \lambda_{Y_{vas}}) = (0,25, 1,50, 0,25, 0,50, 0,25, 0, 0,25, 2,00)$

Para o exemplo, foram simuladas 10 partidas para o confronto final entre Flamengo e Botafogo. Os resultados das partidas são mostrados na tabela abaixo.

Tabela 4: Resultados das simulações da final do campeonato fictício

Mandante	X	Y	Visitante
Flamengo	0	0	Botagogo
Flamengo	1	0	Botagogo
Flamengo	0	0	Botagogo
Flamengo	1	0	Botagogo
Flamengo	1	0	Botagogo
Flamengo	2	0	Botagogo
Flamengo	2	0	Botagogo
Flamengo	1	0	Botagogo
Flamengo	1	0	Botagogo
Flamengo	1	0	Botagogo

3 Análise dos Resultados

O objetivo do estudo é desenvolver um modelo capaz prever qual o resultado das partidas do segundo turno do Campeonato Brasileiro e, dessa maneira, informar o time com maior pontuação ao final do torneio e definir se o Flamengo era uma das equipes favoritas para ser campeã do brasileirão de 2009.

Para isso, em todos os campeonatos utilizados no trabalho, foram realizadas um total de 100 simulações do segundo turno de cada campeonato e, através delas, descobriu-se quais os times que tinham a maior probabilidade ser o campeão.

3.1 Campeonato Brasileiro de 2009

O Campeonato Brasileiro de 2009 foi formado por 380 partidas, divididas em 2 turnos, e um total de 20 equipes participantes. A escolha desse ano como principal objeto de análise se deve ao fato de algumas polêmicas terem sidos apresentadas na sua reta final.

Antes de expor quais foram os resultados controversos se faz necessário contextualizar o que aconteceu naquele ano. Abaixo é mostrado um gráfico com o desempenho, rodada a rodada, dos 5 times que chegaram ao final do campeonato com chances de ganhar o título.

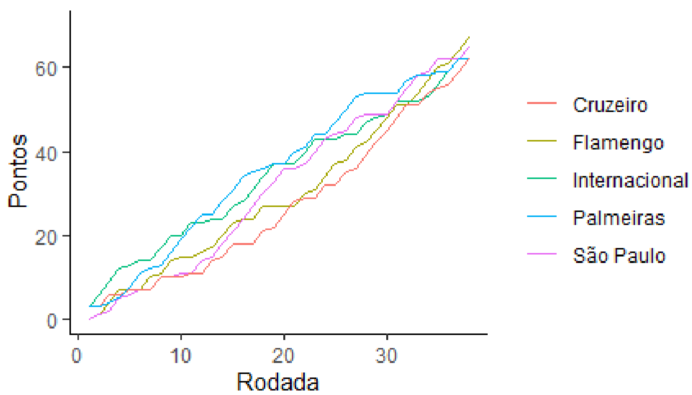

Figura 2: Gráfico da pontuação por rodada

Como são utilizada apenas as partidas jogadas no primeiro turno para ajustar o modelo, alguns pontos merecem ser destacados. Ao fim da 19ª rodada, o campeonato era liderado pela equipe do Internacional (37 pontos), a equipe com mais gols marcados era o time do Barueri (38 gols feitos) e a equipe com menos gols sofridos era o Palmeiras (17 gols sofridos).

Até o fim do 1º turno, os números que o time do Flamengo apresentava não eram bons. Ele estava somente na décima posição com 27 pontos, 10 pontos atrás do líder Internacional, era também o décimo melhor ataque (28 gols feitos) e o sexto time que mais tomou gols (31 gols sofridos).

Baseando-se pela figura 2 e no desempenho até a metade do campeonato, parece haver indícios para acreditar que o Flamengo não era uma das equipes favoritas para ser a vencedora. Somado aos números, existem as polêmicas nos dois últimos jogos, contra o Corinthians e Grêmio, que são rivais diretos dos times que vinham disputando com o Flamengo. Muito se fala, do pênalti que o goleiro do Corinthins não pulou para tentar defender a bola e do Grêmio ter entrado em campo com um time reserva para disputar o último jogo. A tabela 5 mostra a classificação do campeonato ao fim da 19ª rodada.

Tabela 5: Classificação da 19ª rodada

Time	Pontos	GP
Internacional	37	37
Palmeiras	37	31
Goiás	35	37
São Paulo	33	28
Atlético-MG	32	32
Avaí	30	28
Barueri	28	38
Corinthians	28	23
Grêmio	28	33
Flamengo	27	28
Santos	26	34
Vitória	25	27
Atlético-PR	24	21
Cruzeiro	22	19
Botafogo	21	26
Coritiba	19	26
Náutico	18	22
Santo André	18	21
Fluminense	15	21
Sport	15	24

3.1.1 Simulando o Campeonato Brasileiro de 2009

Como apresentado na subseção 2.2.3.1, foram desenvolvidos dois modelos de regressão linear múltipla, um para a variável U e outro para a variável V. A execução dos ajustes dos modelos é elaborada de forma separada para cada ano dos torneios, isto é, se faz necessário ajustar um novo modelo a cada temporada.

Das regressões obtidas através das equações 2.8 e 2.9 os ajustes dos modelos são feitos pela estimação por mínimos quadrados e possuem a finalidade de determinar os coeficientes $\underline{\alpha}$ e$\underline{\eta}$. Utilizando as informações sobre as 19 rodadas disputadas até o fim do primeiro turno, são encontrados os coeficientes e, por meio deles, encontram-se os valores dos estimadores $\hat{\lambda_X}$ e $\hat{\lambda_Y}$ de cada equipe participante.

A partir dos estimadores encontrados é então possível simular, quantas vezes se achar oportuno, os resultados das partidas do segundo turno de cada campeonato. Mediante esses números, podem ser obtidos quantas vezes cada time foi campeão daquele torneio. Para o trabalho aqui feito, foram efetuadas um total de 100 simulações das patidas.

Tabela 6: Tabela de parâmetros

Times	$\hat{\lambda_x}$	$\hat{\lambda_y}$
Atlético-MG	0.82129724	0.79283056
Atlético-PR	0.45932218	0.62582672
Avaí	0.80433304	0.13645757
Barueri	1.73262364	1.24328204
Botafogo	0.44130951	1.01078650
Corinthians	1.12020723	0.56208350
Coritiba	0.86848696	0.80198242
Cruzeiro	0.04672992	1.07519659
Flamengo	0.87831759	1.65357118
Fluminense	0.20144002	1.02458492
Goiás	0.78580252	0.60578515
Grêmio	1.81903575	0.50250652
Internacional	1.67732519	0.89385443
Náutico	0.19786698	1.69312895
Palmeiras	0.68447828	0.09727443
Santo André	0.08661184	0.67595344
Santos	1.22625215	1.01345600
São Paulo	1.23819877	0.19632250
Sport	0.84530144	1.45003947
Vitória	1.28128196	1.16129934

Empregando os parâmetros mostrados pela tabela 6 é possível determinar os resultados das partidas, ou seja, apontar se houve um time vitorioso ou se o jogo terminou empatado. Em posse dessa informação, é viável especificar quem foi o campeão do torneio naquele ano.

Assim, abaixo são apresentadas duas tabelas que expõem quais foram as pontuações finais reais obtidos equipes comparados com a pontuação média encontrada pelas simulações e quais foram as frequências dos clubes que ganharam o título em pelo menos um dos 100 campeonatos simulados.

Tabela 7: Resultados das simulações de 2009

Times	Quantidade de campeonatos	Proporção (%)
Internacional	63	63
Barueri	18	18
Atlético-MG	6	6
Goiás	4	4
Grêmio	4	4
Vitória	2	2
Flamengo	1	1
São Paulo	1	1
Palmeiras	1	1

Tabela 8: Comparação entre realidade e simulações em 2009

Posição	Times	Pontos (real)	Média de pontos (simulações)
1	Flamengo	67	58
2	Internacional	65	68
3	São Paulo	65	58
4	Cruzeiro	62	41
5	Palmeiras	62	55
6	Avaí	57	48
7	Atlético-MG	56	59
8	Grêmio	55	59
9	Goiás	55	59
10	Corinthians	52	51
11	Barueri	49	63
12	Santos	49	54
13	Vitória	48	57
14	Atlético-PR	48	44
15	Botafogo	47	44
16	Fluminense	46	37
17	Coritiba	45	46
18	Santo André	41	36
19	Náutico	38	44
20	Sport	31	42

Pelo o que foi ilustrado na tabela 7, as simulações mostram que o time com a maior chance de se consagrar campeão naquele ano era a equipe do Internacional, já que somou mais pontos que seus concorrentes em 63 das 100 simulações feitas.

Um ponto interessante é a aparição do Barueri como candidato ao título, ganhando 18 vezes nas simulações, um número muito maior do que o dos times que realmente apareceram na disputa. Entretanto, olhando a campanha do time até o fim do 1º turno é possível ver que realmente ele parecia ter chances. Nesse momento em questão, a equipe possuia 28 pontos, estando à frente do Flamengo e Cruzeiro e tinha o melhor ataque do campeonato (38 gols marcados).

Na tabela 3.1.1 é ilustrada uma comparação entre como ficou a classificação final do Campeonato Brasileiro de 2009 e a pontuação média obtida por cada time nas simulações geradas pelo modelo. Nota-se que, para a grande maioria dos casos, o modelo apresenta um resultado bem próximo do que aconteceu na realidade.

Visto que o futebol não é uma ciência exata e é um esporte muito dinâmico, existem outros diversos fatores que tem a capacidade de influenciar os resultados finais de um torneio. Por exemplo, existem as janelas de transferências de jogadores na metade do ano o que implica na perda de material humano e, geralmente, os times perdem os seus melhores atletas nesse período.

Um outro fator muito complicado de se mensurar é o apoio da torcida, não somente nos estádios. Quando um time tem o suporte de sua torcida dentro e fora de campo, até as situações com menores possibilidades podem ocorrer.

Assim, dado as inúmeras variáveis que influenciam o resultado de uma partida de futebol, é possível acreditar que existem evidências para afirmar que modelo é capaz de prever os ganhadores das partidas de um jogo de futebol com um certo grau de acertividade e que o Flamengo, infelizmente, não era o time favorito para ser consagrado campeão em 2009.

3.2 Campeonatos Brasileiros de 2005 a 2008

Afim de analisar se o modelo realmente tem a capacidade de prever o ganhador das partidas de um campeonato com alguma precisão e, consequentemente, o campeão daquele torneio, foram realizadas simulações do segundo turno dos campeonatos de 2005 à 2008. Para isso, será utilizado a mesma quantidade de simulações de 2009 para todos os outros brasileiros.

Buscando facilitar a compreensão e entendimento das infromações sobre os campeonatos, serão mostrados, de maneira separada e de forma decrescente na linha do tempo, os dados encontrados nas simulações. Portanto as primeiras informações mostrada são do Campeonato Brasileiro de 2008, depois 2007, 2006 e 2005.

Tabela 9: Comparação entre realidade e simulações em 2008

Posição	Times	Pontos (real)	Média de pontos (simulações)
1	São Paulo	75	59
2	Grêmio	72	55
3	Cruzeiro	67	60
4	Palmeiras	65	62
5	Flamengo	64	51
6	Internacional	54	44
7	Botafogo	53	57
8	Goiás	53	47
9	Coritiba	53	54
10	Vitória	52	58
11	Sport	52	52
12	Atlético-MG	48	54
13	Atlético-PR	45	40
14	Fluminense	45	40
15	Santos	45	44
16	Náutico	44	44
17	Figueirense	44	52
18	Vasco da Gama	40	57
19	Portuguesa	38	57
20	Ipatinga	35	43

Discutindo um pouco sobre os resultados obtidos para o ano de 2008 é possível notar que o modelo não acertou o campeão mais uma vez, porém é perceptível que o modelo parece acertar muitos resultados das partidas. A diferença média entre a pontuação real e as médais das pontuações das simulações é de apenas 6,95 pontos, o que pode ser dito como duas vitórias e um empate.

Um ponto que chama atenção é a diferença que aparece para os extremos da tabela. A média para o erro entre os valores reais e dos obtidos considerando apenas os dois primeiros últimos colocados é de 17,25 pontos, o que indica uma maior discrepância entre realidade e simulação para essas posições.

Tabela 10: Campeões das simulações para 2008

Equipe	Quantidade de campeonatos
Palmeiras	25
São Paulo	10

A tabela 10 mostra o Palmeiras como o time que tinha mais chances, ganhando em

25 das 100 simulações feitas. Já o São Paulo, o verdadeiro campeão, ganhou em somente 10 ocasiões.

Tabela 11: Comparação entre realidade e simulações em 2007

Posição	Times	Pontos (real)	Média de pontos (simulações)
1	São Paulo	77	56
2	Santos	62	56
3	Flamengo	61	58
4	Fluminense	61	44
5	Cruzeiro	60	65
6	Grêmio	58	47
7	Palmeiras	58	46
8	Atlético-MG	55	53
9	Botafogo	55	66
10	Vasco da Gama	54	65
11	Internacional	54	51
12	Atlético-PR	54	47
13	Figueirense	53	54
14	Sport	51	61
15	Náutico	49	44
16	Goiás	45	55
17	Corinthians	44	44
18	Juventude	41	46
19	Paraná	41	41
20	América	17	35

Para o Brasileiro de 2007, é visto uma diferença um pouco maior do que a do ano de 2008. O valor médio encontrado para as diferenças entre pontuações é de 7,65 pontos, entretanto as maiores discrepância estão entre o primeiro, terceiro e último colocados, com uma diferença média de 18,67 pontos.

Tabela 12: Campeões das simulações para 2007

Equipe	Quantidade de campeonatos
Cruzeiro	29
São Paulo	3

Como mostrado na tabela 12,em 2007 o modelo mostrou um pouco mais de dificuldade para acertar os resultados das partidas. Ele declarou o Cruzeiro como o time com mais chances de ser campeão, enquanto o verdadeiro campeão foi uma das equipes que menos ganhou nas simulações.

Tabela 13: Comparação entre realidade e simulações em 2006

Posição	Times	Pontos (real)	Média de pontos (simulações)
1	São Paulo	78	64
2	Internacional	69	56
3	Grêmio	67	61
4	Santos	64	55
5	Paraná	60	60
6	Vasco da Gama	59	54
7	Figueirense	57	55
8	Goiás	55	44
9	Corinthians	53	41
10	Cruzeiro	53	51
11	Flamengo	52	41
12	Botafogo	51	41
13	Atlético-PR	48	44
14	Juventude	47	47
15	Fluminense	45	57
16	Palmeiras	44	58
17	Ponte Preta	39	55
18	Fortaleza	38	38
19	São Caetano	36	49
20	Santa Cruz	28	47

No ano de 2006 o modelo teve, mais uma vez, um acréscimo na diferença média entre os pontos obtidos de verdade e a média das pontuações dos campeonatos simulados. O erro médio das diferenças ficou em 8,83 pontos. Entretanto, nesse Brasileiro, o modelo acertou o time campeão favorito ao título daquele ano.

Tabela 14: Campeões das simulações para 2006

Equipe	Quantidade de campeonatos
São Paulo	37
Paraná	21
Grêmio	16

Como o modelo informou com acertividade o campeão, a tabela 14 mostra os três times com maiores quantidades de campeonatos ganhos nas simulações. Algo importante e que vale ser destacado é o fato dele ter mostrado outros dois times que terminaram o torneio bem colocados, o time do Grêmio ficou em terceiro e o Paraná em quinto.

Tabela 15: Comparação entre realidade e simulações em 2005

Posição	Times	Pontos (real)	Média de pontos (simulações)
1	Corinthians	81	74
2	Internacional	78	57
3	Goiás	74	59
4	Palmeiras	70	65
5	Fluminense	68	61
6	Atlético-PR	61	56
7	Paraná	61	55
8	Cruzeiro	60	64
9	Botafogo	59	60
10	Santos	59	63
11	São Paulo	58	50
12	Vasco da Gama	56	64
13	Fortaleza	55	57
14	Juventude	55	55
15	Flamengo	55	49
16	Figueirense	53	46
17	São Caetano	52	54
18	Ponte Preta	51	64
19	Coritiba	49	53
20	Atlético-MG	47	44
21	Paysandu	41	53
22	Brasiliense	41	46

No Campeonato Brasileiro de 2005, o modelo indicou uma melhora na previsão dos resultados das partidas. A diferença média entre a pontuação real e a média dos pontos atingidos nas simulações foi de 7,03 pontos.

Tabela 16: Campeões das simulações para 2005

Equipe	Quantidade de campeonatos
Corinthians	59
Ponte Preta	14
Vasco	8

Pela tabela 16 é visto que o modelo apontou corretamente em 59 vezes das 100 simulações que o time campeão do brasileirão seria o Corinthians, o que aconteceu na realidade. Todavia, um ponto que precisa de um cuidado está no fato dos dois outros times que mais ganharam nas simulações não terem brigado pelo título e a equipe da Ponte Preta terminou na parte de baixo da tabela.

Tabela 17: Comparação dos resultados

Ano	Mais vezes campeão (Simulação)	Campeão (real)
2005	Corinthians	Corinthians
2006	São Paulo	São Paulo
2007	Cruzeiro	São Paulo
2008	Palmeiras	São Paulo
2009	Internacional	Flamengo

Portanto, comparando os resultados evidenciados pela tabela 17 e os dados apresentados nas tabelas é possível constatar que o modelo parece conseguir prever o resultado de uma partida com alguma precisão. Ele apontou corretamenteo campeão em 40 % dos casos e, na maioria dos casos em que errou, apontou um time que brigou pelo título daquele ano. Um outro ponto que merece destaque é o de que a pontuação média das simulações não diferir, em tantos pontos, do valor real.

Entretanto, ainda existem algum casos em que a diferença é muito grande e isso pode ser devido a algum fator que não foi levado em consideração no modelo, como a perda de jogadores importantes durante a janela de transferência do meio do ano ou problemas internos dentro do clube.

4 Conclusões

Fazendo uso do método da Soma e Diferença 0 (SD0) foi desenvolvido um modelo que busca predizer quais foram os resultados finais de uma dada partida, isto é, determinar qual fora o placar do jogo e se algum time saiu vitorioso ou se ocorreu um empate. Além disso, era esperado mostrar que a equipe do Flamengo era um dos times favoritos para disputar a briga pelo título de campeão brasileiro de 2009.

Em vista do que foi apresentado durante o trabalho, foi possível verificar que o modelo acerta muito dos resultados das partidas, apontado bem quais são os times vitoriosos, perdedores e quando há empate. Isso pode ser dito porque por mais que o modelo não acerte todos os placares a diferença de pontos entre o que aconteceu na realidade e a média de pontos obtidas pelas simulações não é tão grande para a maioria dos casos. Entretanto, como o futebol é um esporte imprevisível e repleto elementos aleatórios, não era esperado desenvolver um modelo que tivesse a capacidade de prever todos os placares corretamente utilizando apenas dados sobre mando de campo e número de gols.

Por fim, a elaboração da monografia não foi uma das tarefas mais simples, ainda que exista uma quantidade considerável de materiais para fundamentar e traçar o caminho a seguir, escolher qual a melhor rumo a tomar foi complicado. Outra dificuldade encontrada foi a de aplicar o conhecimento encontrado nos materiais utilizados, muitas das vezes os autores assumem que o assunto é trivial e acabam não se aprofundando em alguns conceitos que poderiam ajudar pessoas com menor grau de conhecimento sobre o tema. Todavia, apesar dos percalços encontrados no decorrer da formação do trabalho, foi muito aprazível conseguir criar um modelo de previsão de partidas de futebol.

Referências

ALVES, M. I. F. Inversa generalizada de moore-penrose e solução de norma mínima em delineamentos experimentais. 1990. Disponível em: ⟨https://teses.usp.br/teses/disponiveis/11/11134/tde-20191108-110711/publico/AlvesMariaIzalinaFerreira.pdf⟩.

ARRUDA, M. L. de. *Poisson, Bayes, Futebol e DeFinetti*. Dissertação (Mestrado), São Paulo, Brasil, 2000.

KOKENY, G. História das apostas esportivas. Brasil, 2021. Disponível em: ⟨https://apostasbrazil.com.br/historia-das-apostas-esportivas/⟩.

MONTGOMERY, D.; PECK, E.; VINING, G. *Introduction to Linear Regression Analysis*. [S.l.: s.n.], 2013.

SAMEJIMA, K. Matd41 - introdução aos modelos lineares. 2020.

SUZUKI, A. K.; TAVARES, L. Modelagem estatística para previsão esportiva: Uma aplicação no futebol. São Paulo, Brasil, 2015.

VALERIANO, M. L. História do futebol. Brasil, 2020. Disponível em: ⟨https://escolaeducacao.com.br/historia-do-futebol/⟩.