CLEIBSON APARECIDO DE ALMEIDA

MODELAGEM ESTATÍSTICA PARA PREVISÃO DE RESULTADOS EM JOGOS DE FUTEBOL ON-LINE

Curitiba

Junho/2006

CLEIBSON APARECIDO DE ALMEIDA

MODELAGEM ESTATÍSTICA PARA PREVISÃO DE RESULTADOS EM JOGOS DE FUTEBOL ON-LINE

Trabalho apresentado à disciplina de Laboratório de Estatística I do curso de graduação em Estatística, da Universidade Federal do Paraná.

Orientador: Prof. Msc. Leonardo Bastos

Curitiba

Junho/2006

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

MODELAGEM ESTATÍSTICA PARA PREVISÃO DE RESULTADOS EM JOGOS DE FUTEBOL ON-LINE

Junho/2006
CLEIBSON APARECIDO DE ALMEIDA

RESUMO

Os dados analisados neste relatório fazem parte de um estudo realizado, com o objetivo de criar um modelo estatístico para fazer previsões em jogos de futebol on-line utilizando a metodologia da Análise de regressão. Para um melhor entendimento das variáveis também foi realizada uma Análise Estatística Descritiva com o objetivo de verificar o comportamento dos dados para posterior modelagem e também facilitar a leitura inicial dos dados.

O conjunto de dados foi coletado do jogo de futebol on-line “hattrick” por meio de softwares complementares com funcionalidades de registrar, gravar e tabular os resultados de todas as partidas realizadas. A escolha do hattrick deve-se ao fato de existir em torno de 800000 usuários em todo o planeta e também por ser o jogo da sua categoria que mais se aproxima a uma partida de futebol real.

Após a modelagem foi possível fazer previsões de jogos que não pertenciam à base de dados iniciais e os resultados obtidos com as previsões realizadas foram favoráveis ao uso do modelo de regressão selecionado.

Palavras-chave: Futebol Virtual, Dados Longitudinais, Modelagem Estatística.

1 INTRODUÇÃO

O Hattrick[1] é um jogo de computador simulado via internet que visa à gestão de um clube de futebol. Originário da Suécia, o jogo teve seu inicio em 1995 e conta atualmente com mais de 800000 usuários em todo o mundo. Existem atualmente 102 paises e suas respectivas ligas nacionais. Uma comunidade mundial foi construída em torno do jogo, que coleta informações estatísticas e fala sobre hattrick em conferências.

Depois do registro, o usuário recebe uma equipe. Esse processo pode demorar dependendo do número de usuários que estão numa lista de espera. O usuário novato recebe uma equipe que estava sem usuário das duas últimas divisões do país onde se inscreveu. Para, além disso, recebe um plantel de jogadores, um treinador, um estádio pequeno e uma pequena quantia de dinheiro virtual. Depois disso, o usuário começa a gerir a sua equipe envolvendo contratações e demissões de jogadores, definição das táticas futebolísticas, contratação e demissão de equipa técnica que é composta por médicos, fisioterapeutas, economistas, relações públicas, etc. Outros fatores importantes envolvem denominar ordens para os jogos, fazer convites para jogos amistosos e principalmente manter torcida e patrocinadores felizes com o clube.

Cada temporada dura 4 meses. Se a equipe obtiver sucesso sobe para a divisão superior, por qualificação direta ou disputando uma repescagem com uma equipe da divisão superior e que esteja em processo de rebaixamento. Este esquema de divisões varia de um país para o outro e pode variar de 6 até 14 divisões.

Existem muitas personalizações que um usuário pode fazer em seu próprio clube, como renomear seu estádio, mudar o tipo de treino, ou até contratar um técnico novo. Mas, como qualquer jogo de gerenciamento, o usuário deve ordenar a seus jogadores em que posição eles deverão jogar, pois cada jogador possuí uma característica específica e isso caracteriza a sua posição durante os jogos.

O Hattrick está evoluindo constantemente e seus desenvolvedores adicionam o cada dia novas funcionalidades ao jogo, melhorando as atuais. Embora este seja sempre um debate entre os usuários devido à ética imposta pelos desenvolvedores do jogo, o Hattrick continua centralizado para evitar parcialidades regionais em seu desenvolvimento.

[1] Hattrick – A homepage do jogo de futebol on-line Hattrick é http://www.hattrick.org

Considerando o número de usuários ativos, os países que lideram ligas no Hattrick são a Espanha, Suécia, Holanda, Bélgica, Argentina e Suíça, embora países como Polônia, Estônia, Finlândia, Romênia e Alemanha, entre outros, estejam em direção das posições no topo. O Hattrick também tem Seleções Nacionais, dois para cada país, uma Sub-20 e uma sem restrição de idade. A cada temporada, uma Copa do Mundo é organizada, uma temporada para Sub-20, outra para o time sênior, e assim por diante. O primeiro país a vencer uma copa internacional sobre a Suécia foi a Romênia. O clube dos Campeões da Copa do Mundo tem poucos membros até agora: Suécia, Romênia e Noruega. A Suécia venceu a última Copa do Mundo, após derrotar a Escócia por 8 a 0 na final, tendo os escoceses se classificando surpreendentemente com seu técnico muito talentoso, e com a falta de sorte em cair contra o time mais forte no fim.

Existem muitos softwares complementares disponíveis para auxiliar o usuário na sua experiência com o jogo, embora muitos usuários debatam se eles diminuem ou não o talento do jogo. Ainda não foi constatado nenhum software ou estudo a fim de debater a questão da previsão de resultados ou algo do gênero.

Neste trabalho temos como objetivo fazer previsões do saldo de gols nas partidas do hattrick. Na seção 2 apresentaremos a descrição dos dados, na seção 3 mostraremos os métodos a serem adotados, na seção 4 veremos os resultados obtidos, na seção 5 faremos à conclusão e na seção 6 poderemos verificar as referências utilizadas para a concretização deste trabalho.

2 DESCRIÇÃO DOS DADOS

Neste tópico serão abordadas as características da base de dados e os procedimentos tomados durante a coleta das informações. Também será feita uma análise estatística descritiva dos dados coletados para o estudo.

2.1 Base de dados

A base de dados foi coletada através de um software complementar ao Hattrick, por meio da tecnologia CHPP "Certified Hattrick Product Provider". Esta tecnologia permite que o usuário conecte-se ao servidor Hattrick com o objetivo de visualizar informações dos jogos passados, não cabendo a ele a modificação dos dados. Foram registradas informações longitudinais do comportamento de 4 clubes (Atlétic Stiletto, Sónaveia, Guerreiros de Seth e Chiliks Lucidus) que participam da sexta divisão no campeonato hattrick, durante um período de 150 dias, com total de 64 observações. As medições foram realizadas semanalmente entre os meses de Agosto de 2005 até Janeiro de 2006. Portanto, trata-se de uma base de dados longitudinais, balanceada e regularmente espaçada.

A tabela 1 apresenta as variáveis utilizadas e uma denotação auxiliar que será adotada a partir deste ponto do trabalho.

Variável	Denotação Auxiliar
Clube	clube
Saldo de gols	saldo
Força do clube de interesse	forcA
Formação tática do clube de interesse	formA
Experiência dos jogadores do clube de interesse	expeA
Força do clube adversário	forcB
Formação tática do clube adversário	formB
Experiência dos jogadores do clube visitante	expeB
Mando de campo	campo

Tabela 1 - Denominação das variáveis durante o estudo

Este estudo está caracterizado por quatro clubes que disputaram a temporada 16, na sexta divisão do campeonato hattrick. Cada temporada é composta por 16 semanas, portanto foram colhidas 16 observações para cada clube totalizando 64 jogos ao todo. O nome dos clubes observados são Atletic Stiletto, Chilics Lucidus, Sónaveia e Guerreiros de Seth. Todos os clubes pertencem a usuários com capacidades e conhecimentos similares sobre o hattrick, isso se torna importante sendo considerado como um controle para o estudo.

Antes de falarmos sobre a força é necessário entender as escalas utilizadas pelo hattrick. Clubes de sexta divisão não passam de excelente, mas é importante esclarecer todos os níveis conforme tabela 2.

Lendária = 20	Utópica = 19	mágica = 18	mítica = 17	colossal = 16
titânica = 15	sobrenatural = 14	genial = 13	magnífica = 12	brilhante = 11
fenomenal = 10	formidável = 9	excelente = 8	boa = 7	razoável = 6
inadequada = 5	fraca = 4	ruim = 3	péssima = 2	terrível = 1

Tabela 2 - Níveis dos jogadores e times do hattrick

É interessante observar que neste trabalho foi utilizado apenas a escala numérica. “A escolha de uma escala numérica facilita a leitura dos dados”. (Li Min Li)

A força do time é baseada nas características individuais de cada jogador pertencente ao plantel daquele clube. Cada jogador tem 8 habilidades básicas, além de alguns fatores adicionais que alteram o seu desempenho em diferentes situações. Vamos detalhar cada uma das habilidades:

Resistência

Controla quanto um jogador perde de seu desempenho durante o segundo tempo da partida. Isto é de grande importância para jogadores de meio-campo, principalmente os centrais, mas também para atacantes defensivos e zagueiros ofensivos, e de alguma importância para outros jogadores também, embora pequena.

Armação

Toda equipe de sucesso precisa de bons armadores, principalmente no meio-campo, garantindo posse de bola e, assim, mais oportunidades de gol. (Penna, 1998)

Ala

Esta é a habilidade de criar chances de gols pelas laterais. É obviamente mais importante para os alas, apesar de laterais com essa habilidade também serem beneficiados. (Penna, 1998)

Finalização

É uma habilidade indispensável para os atacantes. O lugar da bola é no fundo da rede e para isso acontecer precisa-se de atacantes com faro de gol. (Penna, 1998)

Goleiro

Normalmente é a única habilidade que um goleiro consegue adquirir.

Assistências

O artilheiro leva a fama quando o time faz um gol, mas quem faz a assistência é tão importante quanto ele. Esta habilidade deve estar presente em atacantes e meio-campos que tem o papel de armação.

Defesa

É a habilidade de parar ataques do adversário. Inquestionavelmente é a característica mais importante para todos os tipos de jogadores de defesa, sendo que meio-campos também podem precisar, quando forem escalados para ser volantes de contensão. (Penna, 1998)

Bola parada

Um jogador apontado como cobrador de faltas é aquele que baterá todas as faltas e pênaltis do time, e em caso de disputa de pênaltis, ele será o primeiro a cobrar. Importante para todos os jogadores.

A figura 1 ilustra um jogador do clube Chilics Lucidus, perceba que a característica resistência e goleiro indicam que o jogador tem o perfil para jogar a partida como goleiro, caso seja escalado em outra posição o time não terá aproveitamento adequado.

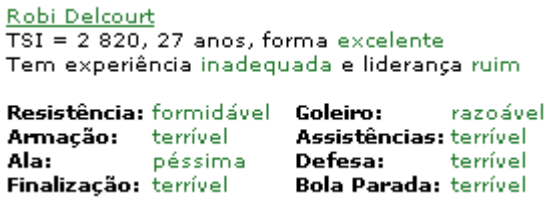


Figura 1 – Características específicas do Jogador

Ainda na figura anterior podemos observar a experiência do jogador. A variável experiência é baseada na soma total das experiências dos jogadores que disputaram determinada partida. Para clubes da sexta divisão este item variam entre 1 e 4. “Atacantes experientes podem marcar gols usando sua experiência. Meio-campos e zagueiros inexperientes podem dar uma chance de gol ao adversário.” (hattrick).

O fator mando de campo indica onde foi realizado determinada partida. “O time da casa é ajudado por seus torcedores. Você normalmente consegue uma posse de bola maior jogando em casa que jogando como visitante. Os árbitros também tendem a marcar mais pênaltis para o time da casa.” (hattrick).

2.2 Estatística Descritiva

Este procedimento tem por finalidade verificar o comportamento das variáveis envolvidas no estudo. A variável saldo de gols é a resposta deste estudo. A tabela 3 mostra as características básicas desta variável. Note que a média é maior que zero indicando que o times estudados saíram de campo com mais vitórias do que derrotas.

Variável	Mínimo	Máximo	Média	Desvio Padrão
saldo de gols	-7	10	2,18	3,93

Tabela 3 – Comportamento do saldo de gols

Ainda observando o saldo de gols vamos verificar visualmente o comportamento dessa variável. Perceba na figura 2 que grande parte das observações estão bem próximos ao valor médio descrito na tabela 3, é um indício de que os dados comportam-se de acordo com a distribuição normal.

Se observarmos o lado direito e esquerdo deste gráfico, podemos ver que existe um "peso visual" tendendo para o lado direito sinalizando que na maioria das vezes o time de interesse venceu seus jogos. Ao verificar o lado esquerdo podemos ver que nas vezes em que o time adversário venceu, a vitória foi por um saldo igual ou maior a 2 gols na maioria dos casos.

Ainda na figura 2 é possível fazer análises comportamentais que não são comuns no futebol real, como por exemplo, o fato de ter poucos empates durante as partidas.

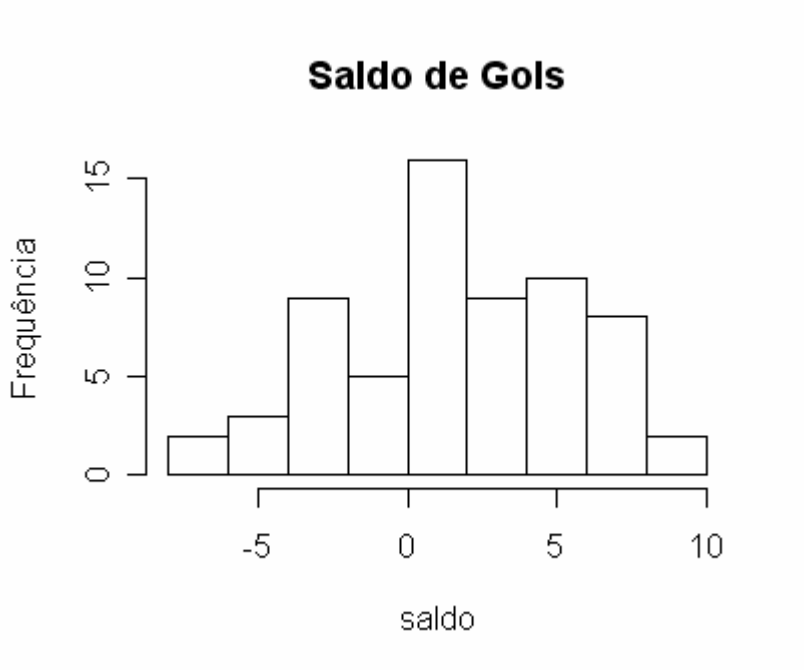


Figura 2 – Distribuição do saldo de gols dos times estudados

Ao analisarmos o saldo de gols isoladamente podemos tomar conclusões precipitadas em relação ao estudo e concluir que o time de interesse sempre irá vencer baseado na figura 2, que ilustra um peso para o lado direito. Essa afirmação poderá ser debatida mais tarde, pois precisamos verificar o comportamento do saldo de gols em relação às demais variáveis do estudo.

Observando a tabela 4, vemos que os times de interesse e adversário obtiveram médias similares indicando que apesar dos limites de máximo e mínimo do time adversário ser maior, o time de interesse conseguiu jogar suas partidas com uma variação de força bem menor que o time adversário. Isso se deve ao fato de que no inicio da temporada as chaves em que os times disputaram seus jogos são constituídas por um clube promovido da sétima divisão que neste caso contribuiu para o valor mínimo da força e outro rebaixado da quinta divisão que como conseqüência contribuiu para o valor de máximo.

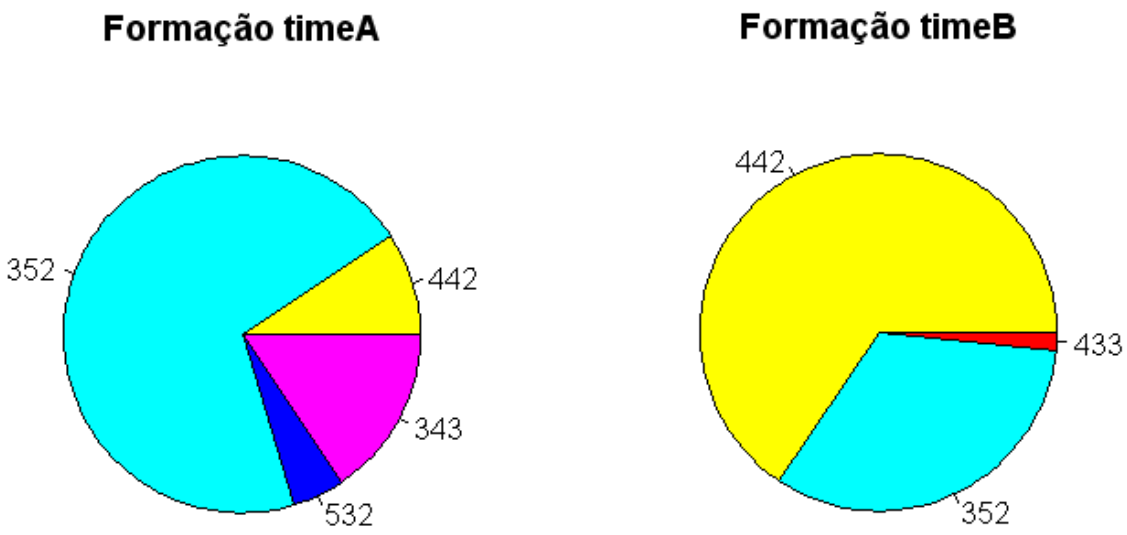
Variável	Mínimo	Máximo	Média	Desvio Padrão
força time de interesse	18	32,5	25,13	3,39
força time adversário	10	35	21,43	6,28

Tabela 4 - Comportamento da força dos times

A figura 3 mostra a formação tática do time de interesse e time adversário. Podemos ver que o time de interesse (figura 3 [a]) preferiu utilizar o esquema tático baseado na formação 352, esta formação transforma laterais em alas ofensivos a fim de deixar o time com maiores chances para criar jogadas de ataque pelas laterais, porém existe um enfraquecimento na defesa do time.

Já o time adversário (figura 3 [b]) teve a maioria de seus jogos utilizando uma formação mais conservadora, o famoso 442. Essa formação tem por base deixar o time com poderes ofensivos sem abrir mão da defesa, certamente é a tática ideal para confrontar times com forte poder ofensivo devido a grande chance dos jogadores da defesa roubarem à bola e criarem jogadas de contra-ataque.

Agora nos resta saber o quanto essas formações irão influenciar no resultado da partida, e isso será visto mais adiante.



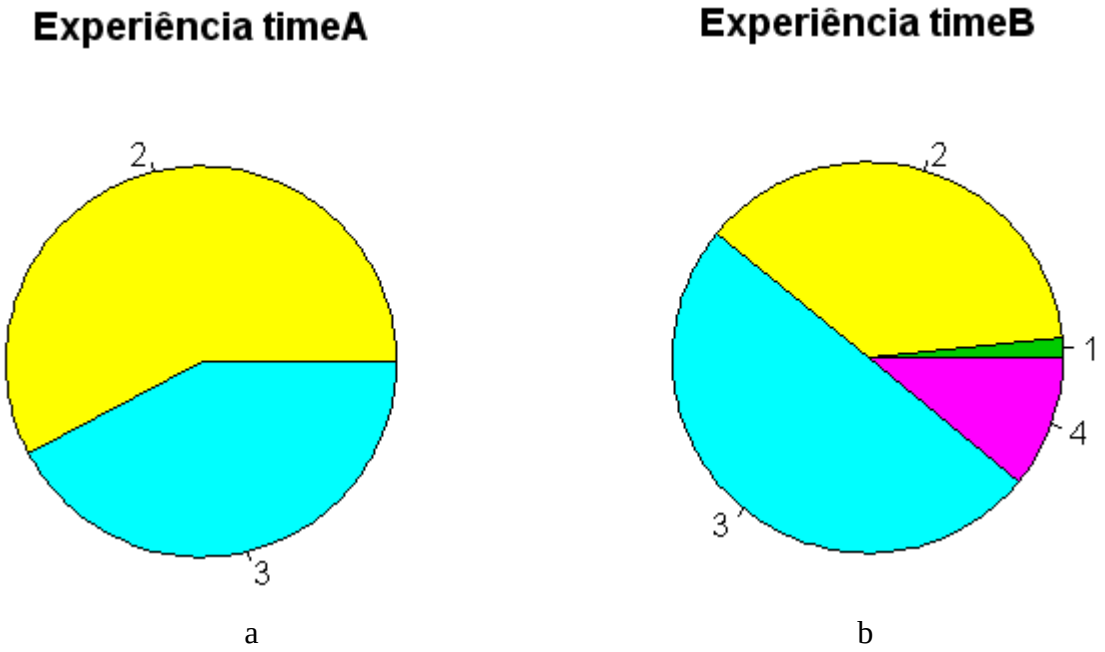
a b

Figura 3 – Formação Tática utilizada pelo time de interesse e adversário

A experiência dos times mostra que houve uma prevalência nos níveis 3 e 4, porém houveram casos em que o time adversário (figura 4 [b]) obteve níveis 1 e 4. Essa variação do time adversário deve-se ao fato que foi comentado anteriormente sobre os clubes que saem rebaixados e promocionados, respectivamente da quinta e sexta divisão. Como cada divisão

têm clubes com um nivelamento parecido, clubes de quinta e sétima divisão tem níveis diferentes quando iniciam um campeonato na sexta divisão.

Ao decorrer do campeonato esse nivelamento é corrigido, deixando esses clubes com níveis de experiência apropriados para a divisão em que estão disputando o campeonato.



a b

Figura 4 – Experiência do time de interesse e adversário

O mando de campo (figura 5) foi equilibrado mesmo não sendo controlado durante o estudo. Isso deve-se ao fato em que durante a temporada o time de interesse confrontou o time adversário em duas ocasiões, sendo uma fora de casa e outra em seu próprio campo. A diferença de quatro partidas em favor dos jogos fora de casa, deve-se ao fato em que o time de interesse foi punido em alguma ocasião e tal punição fez com que esse time perdesse seu mando de campo.

Eventualmente quando o time de interesse jogou fora de casa, o time adversário estava em seu próprio campo.

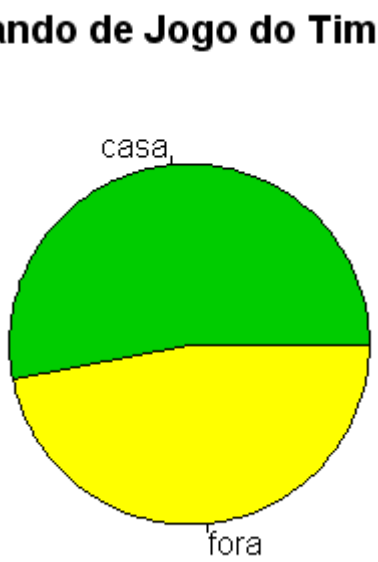


Figura 5 – Mando de jogo relativos ao time de interesse

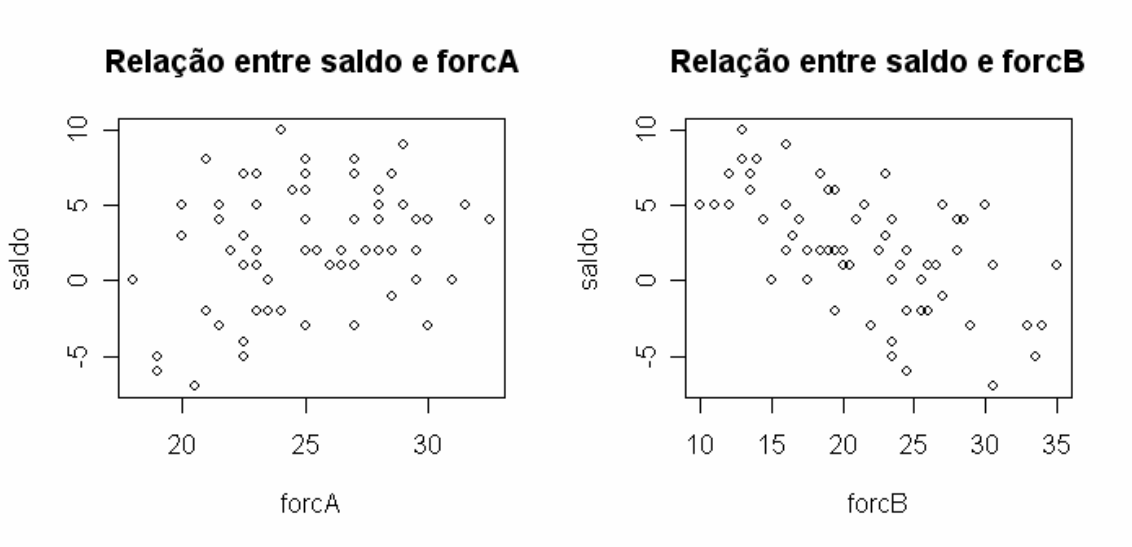
2.3 Relações entre Variáveis

Em um primeiro contato com os dados, será feita uma análise descritiva para avaliar o comportamento das variáveis presentes no estudo. Esta análise tem por finalidade dar rumo aos procedimentos a serem tomados durante a modelagem estatística.

Temos que verificar se o saldo de gols depende da força, formação tática, experiência dos jogadores e mando de campo. Para entender melhor o comportamento do saldo de gols em relação a cada uma dessas variáveis vamos fazer um estudo com olhar clínico em cada caso.

2.3.1 Saldo x Força

A melhor forma para avaliarmos o saldo em relação à força é fazer um gráfico de dispersão entre essas duas variáveis. Com a plotagem do gráfico de relação entre saldo e força foi possível verificar um comportamento linear positivo para o time de interesse, ou seja, conforme aumenta a força do time de interesse também aumenta o saldo de gols. No caso do time adversário a situação foi inversa porque o saldo de gols é baseado no time de interesse explicando a tendência linear negativa mostrada pela figura 6.



a b

Figura 6 – Relação linear entre saldo e força dos times

Para facilitar o entendimento da relação existente entre saldo e força mostrado no gráfico anterior, foi criada uma razão entre forcA e ForcB. A figura 7 mostra a relação entre o saldo e a razão das forças, entendendo que quando esta razão for positiva significa que o time de interesse venceu a partida.

Podemos ver na figura 7 a tendência linear positiva entre o saldo e razão (forcA/forcB). Isso indica que a força do time de interesse ou time adversário está relacionada com o saldo de gols, ou seja, uma razão positiva mostra que o time de interesse venceu o jogo e sua força foi elevada devido à vitória.

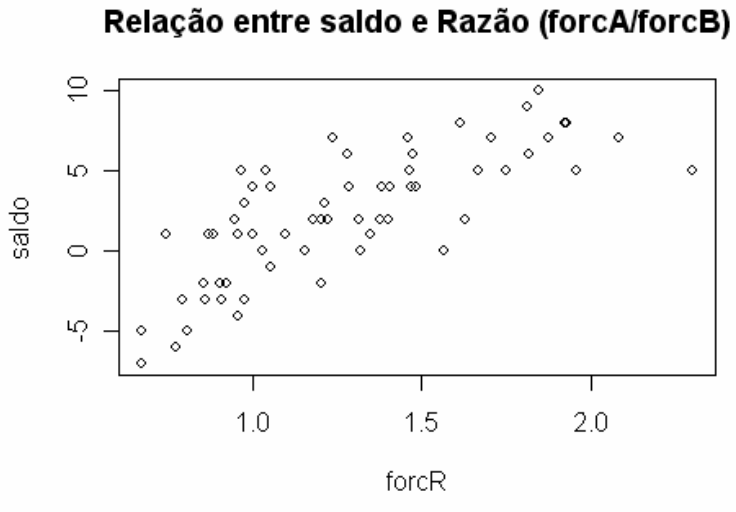


Figura 7 – Relação linear entre saldo e razão(forcA/forcB)

Para averiguar esta relação é melhor observar a tabela 5, onde foi realizado um teste de correlação de Pearson para verificar a correlação entre saldo e força.

O saldo em relação à forcA teve uma correlação de certo ponto duvidosa resultando no valor de 0, 28, porém seu p-valor mostrou-se significativo indicando que saldo e forcA estão realmente correlacionados.

Com p-valor<0,05, podemos rejeitar Ho e dizer que as forças podem explicar o saldo de gols. É interessante observar que a razão entre forcA e forcB nos forneceu um resultado mais conclusivo para correlação e também mais perceptível pela figura 7.

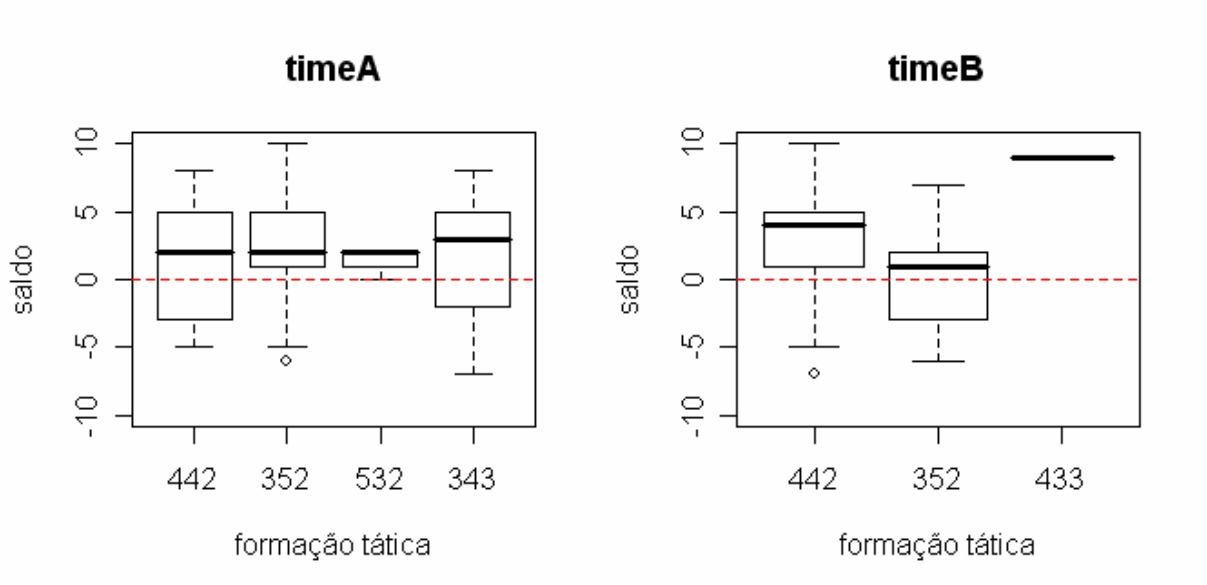
	Saldo x forcA	Saldo x forcB	Saldo x Razão (forcA/forcB)
Correlação	0,28	-0,62	0,76
p-valor	0,021	2,62e-8	1,27e-13

Tabela 5 – Teste de Correlação de Pearson para saldo e força

2.3.2 Saldo x Formação tática

Agora observando o boxplot (figura 8) que mostra a relação do saldo e formação, podemos observar que o time de interesse (figura 8 [a]) manteve sua média de saldo positiva independente da formação tática utilizada durante as partidas, porém os valores extremos (máximo e mínimo) mostram que durante a utilização do modelo tático 442 e 343 o time de interesse teve um comportamento irregular (perdendo e ganhando partidas) e quando utilizou o modelo tático 352 o time manteve uma boa regularidade e ainda com uma média de saldo positivo, mostrando que o time venceu mais de 75% dos jogos quando utilizou esta formação.

Já o time adversário (figura 8 [b]) jogou melhor quando utilizou a tática 442, vencendo mais de 75% de suas partidas nesse esquema. Quando jogou na formação 352, o time B levou desvantagem e perdeu aproximadamente metade dos seus jogos nesta formação. É complicado comentar sobre a tática 433, pois tivemos esta formação em apenas uma partida dentre as 64 jogadas.



a b

Figura 8 – Boxplot do saldo e formação tática dos times

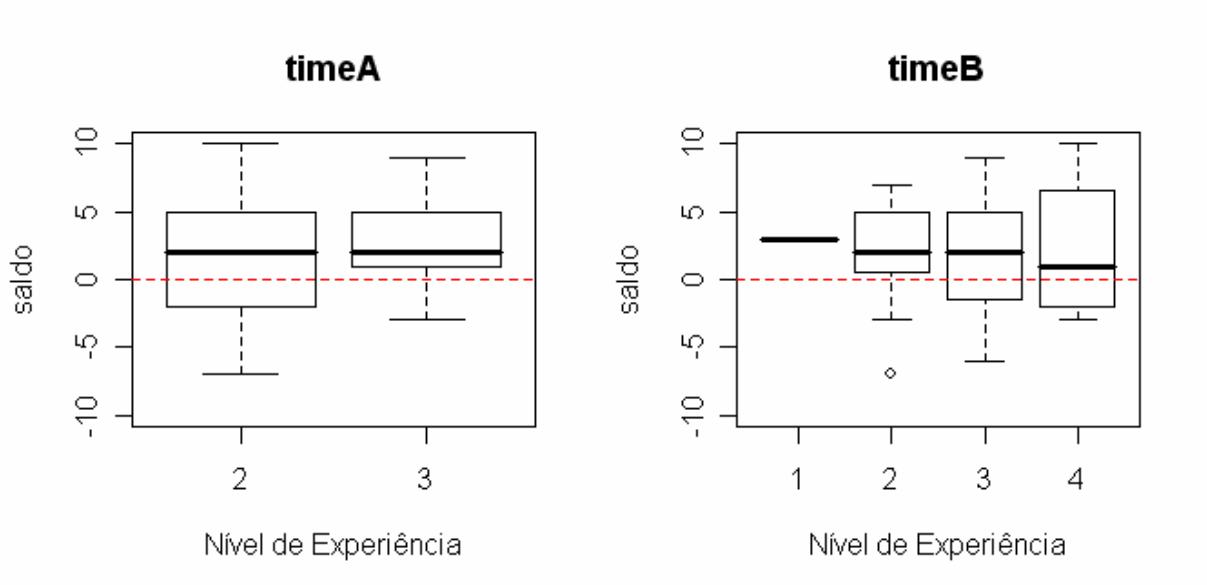
2.3.3 Saldo x Experiência

Outra variável importante no estudo é a experiência dos jogadores, o nível de experiência varia entre 1 e 20. Em nosso caso, que engloba times da 6ª divisão esta amplitude inicia em 1 e vai até 4. Neste estudo obtivemos níveis 2 e 3 para o time de interesse e 1, 2, 3 e 4 para o time adversário.

Ao fazer um boxplot para verificar o comportamento do saldo em relação à experiência (figura 9) tivemos a mesma média de saldo de gols para os dois níveis do time de interesse (figura 9 [a]), porém as valores de máximo e mínimo tiveram diferenças indicando que quando o time de interesse teve experiência igual a 3 o time venceu a maioria de seus jogos.

No caso do time adversário (figura 9 [b]) foi observado quatro níveis de experiência (1, 2, 3 e 4), ressaltando que o nível 1 foi observado em apenas uma ocasião. Como era de se esperar o time adversário teve melhores saldos quando teve uma experiência igual a 3 e 4.

Partindo dessa análise podemos dizer que quanto maior a experiência dos jogadores, maior é chance do time vencer a partida, consequentemente maior saldo de gols a favor do time.



a b

Figura 9 – Boxplot do saldo e Nível de Experiência dos Jogadores

2.3.4 Saldo x Mando de campo

O mando de campo é a última variável a ser comentada, porém não deixa de ser mais ou menos importante que as demais. Sua importância deve-se ao fato de que quando o time joga em seu próprio campo seu saldo é positivo indicando a vitória do time.

Podemos ver este comportamento pela figura 10, que nos mostra um boxplot dos jogos do time de interesse realizado fora e dentro de seu campo. Quando jogou fora teve um saldo inconsistente que variou entre valores negativos e positivos, na maioria das vezes negativos.

Quando jogou dentro de seu campo o time de interesse teve um saldo positivo mostrando indícios de que jogar em casa facilita a vitória. A média do saldo de gols em relação ao mando de campo foram próximas para os jogos dentro e fora de casa indicando que apesar de influenciar o resultado da partida, o mando de campo não influenciou na média de gols do time de interesse.

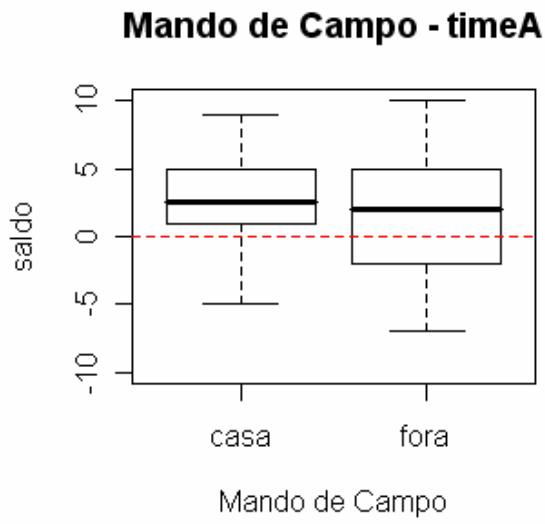


Figura 10 – Boxplot do saldo e Nível de Experiência dos Jogadores

3 MÉTODOS

Nesta seção será relatada toda a metodologia utilizada no trabalho. Tal metodologia tem por finalidade dar embasamento teórico para os resultados obtidos na seção 4.

3.1 Introdução

A Análise de Regressão é, entre as técnicas estatísticas, a mais utilizada na prática quando se deseja modelar o relacionamento entre as variáveis (dependente e independentes). Mais detalhes sobre análise de regressão pode ser obtida em Draper (1971) ou Neter e Wasserman (1974).

A origem do termo "regressão" deu-se pelo estatístico inglês Francis Galton quando estudou o relacionamento das alturas de pais e filhos. As aplicações desta técnica são numerosas e ocorrem em quase todos os campos científicos, sendo que a seguir, é apresentada de forma resumida esta técnica.

3.2 Modelo Linear Geral de Regressão

Seja o modelo:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

onde Y é o vetor aleatório de resposta, B é o vetor de parâmetros de dimensão p, X é a matriz do modelo de ordem n x p e $\underline{\varepsilon}$ é o vetor aleatório de erros de dimensão n. Assim tem-se:

$$\underset{n\,x\,1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} \qquad \underset{n\,x\,p}{X} = \begin{bmatrix} 1 & X_{11} & . & . & X_{1,p-1} \\ 1 & X_{21} & . & . & X_{2,p-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & . & . & X_{n,p-1} \end{bmatrix} \qquad \underset{p\,x\,1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} \quad \underset{n\,x\,1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Admite-se para o modelo as seguintes suposições:

1) o vetor de erros $\underline{\varepsilon}' = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]$ é aleatório, ou seja, as componentes $\varepsilon_i \ i = 1, 2, ..., n$ são variáveis aleatórias;

2) a esperança de cada componente de $\underline{\varepsilon}$ é zero, ou seja, $E(\underline{\varepsilon}) = \underline{0}$; ;

3) as componentes do vetor $\underline{\varepsilon}$ não são correlacionadas ou melhor $(\varepsilon_i, \varepsilon_j) = 0, i \neq j$ e possuem variância constante, σ^2. Assim, a matriz de covariâncias de $\underline{\varepsilon}$ é a matriz diagonal $\sigma^2 I_n$ onde I_n é a matriz identidade de ordem n, $V(\underline{\varepsilon}) = \sigma^2 I_n$.

O modelo (2.72) com as três suposições anteriores é conhecido como Modelo Linear de Gauss Markov e o Teorema de Gauss-Markov garantem que sob as três suposições e com $X'X$ não singular, o estimador não viciado uniformemente de mínima variância (UMVU) do vetor $\underline{\beta}$ e para a variância σ^2 são, respectivamente:

$$\underline{\hat{\beta}} = (X'X)^{-1}(X'\underline{Y})$$

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Uma suposição exigida, além das três já citadas, para o modelo de regressão é a seguinte:

4) a distribuição de $\varepsilon_i \ i = 1,2, ..., n$ é a Normal

.

Considerando esta suposição, tem-se o modelo de Gauss-Markov Normal e

$$Y_i \sim N\left(\sum_{i=1}^{p}\beta_i x_i, \sigma^2\right).$$

3.3 Análise da Variância da Regressão

A Análise da Variância é uma das técnicas estatísticas cujas bases foram lançadas por Fisher. Esta é a técnica geralmente usada para verificar se o ajuste de regressão existe. É

comum construir-se um quadro que resume as informações da Análise de Variância, para um modelo geral:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{p-1} X_{p-1,\,i} + \varepsilon$$

para $p > 2$ parâmetros, tem-se o quadro 2. A seguir.

Quadro 3.1: Análise de variância

Fonte de variação	Soma de quadrados	G.L.	Quadrado médio	F
Regressão	$SQ\text{Regr} = \underline{\hat{\beta}}' X' \underline{Y} - n\bar{y}^2$	$p-1$	$\frac{SQ\,\text{Re}\,gr}{p-1}$	$\frac{SQ\,\text{Re}\,gr}{p-1} / \frac{SQR}{n-p}$
Residual	$SQR = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}' X' \underline{Y}$	$n-\text{p}$	$\frac{SQR}{n-p}$	
Total	$SQT = \underline{Y}'\underline{Y} - n\bar{y}^2$	$n-1$	$\frac{QMT}{n-1}$	

Tabela 6 - Quadro geral para análise de variância

O teste feito com a estatística F (última coluna do quadro 2.1) é o da hipótese nula H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$, ou seja, se existe regressão dos X's para Y, ou melhor, se existe relação linear entre a variável resposta Y e as covariáveis $X_i \;\; i = 1, 2, \ldots, p-1$.

3.4 Verificação dos Pressupostos do Modelo

a) Homocedasticidade

Homocedasticidade é a variância constante dos resíduos. Esta é uma propriedade essencial, que deve ser garantida, sob pena de invalidar toda a análise estatística. deseja-se que os erros sejam aleatórios, ou seja, não devem ser relacionados com as características dos imóveis. Se isto não ocorre, há heterocedasticidade. Significa dizer que as chances de ocorrerem erros grandes (ou pequenos) variam conforme o tipo de imóvel. Há tendências nos erros. As conseqüências da

heterocedasticidade são que as estimativas dos parâmetros da regressão $(\beta_0,\ \beta_1,\ \beta_2,\ \ldots,\ \beta_p)$ não são tendenciosas mas são ineficientes e as estimativas das variâncias são tendenciosas.

Os testes t e F tendem a dar resultados incorretos. Neste caso, os resultados não são confiáveis, ou seja, o modelo pode parecer bom, mas ele não é adequado aos dados, na verdade.

A homocedasticidade pode ser verificada, entre outros, através de gráficos de resíduos (erros). Os gráficos dos erros contra os valores reais e contra os valores calculados pela equação são importantes. Se os pontos estão distribuídos aleatoriamente, sem demonstrar um comportamento definido, há homocedasticidade. Mas se existe alguma tendência (crescimento, decrescimento ou oscilação), então há heterocedasticidade. Havendo heterocedasticidade, podem ser feitas transformações nas variáveis (geralmente logarítmicas)

b)Independência serial dos resíduos (não-autocorrelação)

Existe autocorrelação quando os erros são correlacionados com os valores anteriores ou posteriores na série. Este problema também é chamado de correlação serial. Pode surgir por especificação incorreta do modelo da regressão, por causa de erros na forma do modelo ou por exclusão de variáveis independentes importantes para a análise. Existindo autocorrelação, os estimadores ordinários de mínimos quadrados não são mais os melhores estimadores lineares não-tendenciosos (as variâncias amostrais dos coeficientes estimados para a equação serão excessivamente grandes, essas variâncias serão subestimadas, as fórmulas perderão a validade e serão obtidas predições ineficientes). Neste caso existirão outros métodos que produzem menor variância amostral nos estimadores. Além disso, em presença de correlação serial, os testes de significância (t e F) e de construção de intervalos de confiança dos coeficientes da regressão também oferecem conclusões incorretas, isto é, as regiões de aceitação e os intervalos de confiança podem ser mais largos ou mais estreitos do que os calculados, dependendo da tendência ser positiva ou negativa.

A verificação da autocorrelação pode ser feita pela análise do gráfico dos resíduos cotejados com os valores preditos, onde este deve apresentar pontos dispersos aleatoriamente, sem nenhum padrão definido ou pelo este de Durbin-Watson.

c) Normalidade dos resíduos

A Análise de Regressão baseia-se na hipótese de que os erros seguem uma distribuição Normal (distribuição de Gauss). A condição de normalidade dos resíduos é fundamental para

a definição de intervalos de confiança e testes de significância. Ou seja, em presença de falta de normalidade, os estimadores são não-tendenciosos, mas os testes não têm validade, principalmente em amostras pequenas. Entretanto, pequenas fugas da normalidade não causam grandes problemas.

A não-normalidade dos resíduos pode ser causada por violações de outras condições básicas, tais como a heterocedasticidade (variância não constante dos erros) ou a escolha de um modelo incorreto para a equação.

A verificação da normalidade pode ser feita pelos testes de aderência não-paramétricos, como por exemplo, o de Kolmogorov-Smirnov (Campos, 1983).

d) Outliers

Denomina-se outlier um dado que contém grande resíduo em relação aos demais que compõem a amostra e assim tem comportamento muito diferente dos demais (Dantas, 1998, p.112).

É extremamente importante controlar os outliers, porque em virtude da forma de estimação da equação, um erro grande modifica significativamente os somatórios, alterando os coeficientes da equação. Assim, um saldo apenas pode modificar a equação.

Não existem limites fixos, mas geralmente se adota o intervalo de 2 desvios-padrão em torno da média dos erros. Como a média tem de ser zero, os resíduos padronizados

$$\left(\frac{\varepsilon_i}{dp_y}\right)$$

devem estar entre 3 e -3. Os valores de saldo com erros que ultrapassam estes limites são elementos suspeitos e devem ser analisados cuidadosamente. A existência de outliers deve sempre ser interpretada como um sinal de problemas na amostra.

e) Colinearidade ou multicolinearidade

Define-se multicolinearidade como sendo o problema geral, a existência de relações lineares entre as variáveis independentes, de tal forma correlacionadas umas às outras,

tornando-se difícil, se não impossível isolar suas influências separadas e obter uma estimativa precisa de seus efeitos relativos (Johnston, 1986). Quando a relação é exata tem-se o caso da multicolinearidade perfeita. Na prática atual, raramente encontram-se variáveis independentes que são perfeitamente relacionadas. Esse caso não traz problemas, pois é facilmente detectado e pode ser resolvido simplesmente eliminando uma ou mais variáveis independentes do modelo. O interesse no que se refere à multicolinearidade está nos casos em que ela ocorre com alto grau, isto é, quando duas variáveis independentes estão significativamente correlacionadas ou quando há uma combinação linear entre um conjunto de variáveis independentes. Assim, a multicolinearidade é mais uma questão de grau do que de natureza (Elian, 1988).

O fato de muitas funções e regressões diferentes proporcionarem bons ajustes para um mesmo conjunto de dados é porque os coeficientes de regressão atendem várias amostras onde as variáveis independentes são altamente correlacionadas. "Assim, os coeficientes de regressão estimados variam de uma amostra para outra quando as variáveis independentes estão altamente correlacionadas. Isso leva a informações imprecisas a respeito dos coeficientes verdadeiros" (Neter e Wasserman, 1974, p.344).

A multicolinearidade geralmente é causada pela própria natureza dos dados, principalmente nas áreas de economia com variáveis que representam valores de mercado. Algumas vezes a multicolinearidade pode também ocorrer devido à amostragem inadequada (Elian, 1998).

Em Análise de Regressão Linear Múltipla, existe um freqüente interesse com relação à natureza e significância das relações entre as variáveis independentes e a variável dependente. "Em muitas aplicações de administração e economia, freqüentemente encontram-se variáveis independentes que estão correlacionadas entre elas mesmas e, também, com outras variáveis que não estão incluídas no modelo, mas estão relacionadas à variável dependente" (Neter e Wasserman, 1974, p.339).

A existência de multicolinearidade tendo sido detectada e considerada prejudicial indica que o pesquisador deve procurar soluções para suavizar seus efeitos ruins. Várias medidas corretivas têm sido respostas, desde simples às mais complexas, para suavizar os efeitos provocados pela multicolinearidade (Elian, 1988, p.131-134).

Algumas soluções para o problema de multicolinearidade são através de: remoção de variáveis, ampliação do tamanho da amostra, adoção de técnicas estatísticas como Análise de Componentes Principais, entre outras.

3.5 Poder de Explicação do Modelo

Para se medir o quanto a variabilidade total dos dados é explicada pelo modelo de regressão, compara-se a Soma de Quadrados da Regressão com a Soma de quadrados Total e tem-se o coeficiente de determinação ou de correlação múltipla ao quadrado R^2,

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad 0 < R^2 < 1$$

Quando o ajuste é bom o modelo explica boa parte da variação total e conseqüentemente o valor de R^2 é próximo de 1. O coeficiente de determinação é uma medida da qualidade do ajuste.

3.6 Relações entre Variáveis

O coeficiente de correlação é uma medida estatística importante na análise em um modelo de regressão. O grau de relação entre as variáveis, que expressa quão bem essas variáveis estão relacionadas entre si é definido numericamente pelo Coeficiente de Correlação, parâmetro representado por ρ. Com base em n observações do par(X,Y) este parâmetro é estimado pela estatística,

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(X_i - \overline{X})^2 (Y_i - \overline{Y})^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Onde:

$\overline{X}$ é a média da variável independente X;

$\overline{Y}$ é a média da variável dependente Y;

σ_{xy} é a covariância amostral de X e Y;

σ_x é o desvio padrão de X;

é o desvio padrão de Y.

O coeficiente de correlação varia entre os limites –1 e 1 podendo, portanto, ser positivo ou negativo $(-1 \leq \rho \leq 1)$. Quando o coeficiente de correlação é nulo $(\rho = 0)$, significa que não existe nenhum relacionamento entre as variáveis. E quando o coeficiente de correlação é igual à unidade, -1 ou +1, tem-se um relacionamento perfeito entre elas. O sinal (+) ou (-) das variáveis indica a relação direta ou indireta existente entre as variáveis. O grau de relacionamento entre as variáveis, definido numericamente pelo valor $\hat{\rho}$, em Pereira (1970), pode ser assim interpretado:

Coeficiente		**Correlação**
$\lvert\hat{\rho}\rvert = 0$	.....	relação nula
$0 < \lvert\hat{\rho}\rvert \leq 0{,}30$	.....	relação fraca
$0{,}30 < \lvert\hat{\rho}\rvert \leq 0{,}70$	.....	relação média
$0{,}70 < \lvert\hat{\rho}\rvert \leq 0{,}90$	.....	relação forte
$0{,}90 < \lvert\hat{\rho}\rvert \leq 0{,}99$	.....	relação fortíssima
$\lvert\hat{\rho}\rvert = 1$	.....	relação perfeita

Deve ser observado também que nem sempre uma elevada correlação entre duas variáveis representa a existência de relação de causa e efeito entre as mesmas; é necessário analisar se a correlação é absurda. Esses casos dão origem às chamadas de influência no caso.

O estudo do relacionamento entre um conjunto de variáveis pode ser realizado aplicando diversas técnicas, desde os coeficientes de correlação de Pearson, de Spearman, Análise Fatorial e a Análise de Componentes Principais.

A estatística 2.7.7 é conhecida como o coeficiente de correlação linear de Pearson e é uma medida usada no estudo da relação linear existente entre duas variáveis X e Y.

3.7 Seleção de Variáveis Regressoras

Um dos problemas mais freqüentes em Análise de Regressão é a seleção do conjunto de variáveis independentes a serem incluídas no modelo (Neter e Wasserman, 1974, p.371).

O pesquisador deve especificar o conjunto de variáveis independentes a ser empregado para descrever, controlar ou predizer a variável dependente. Um problema muito difícil de relacionamento que aparece na seleção de variáveis é quando uma equação de regressão é construída com o objetivo de predição e envolve muitas variáveis. Talvez, muitas delas contribuam
pouco ou nada para precisão da predição. A escolha apropriada de algumas delas fornece a melhor predição, porém quais e quantas devem ser selecionadas? (Snedecor e Cochram, 1972, p.412-413).

Em algumas áreas, a teoria pode ajudar na seleção das variáveis independentes a serem empregadas e na especificação da forma funcional da relação de regressão. em tais áreas, os experimentos podem ser controlados para fornecer dados sobre a base de que os parâmetros de regressão podem ser estimados e a forma teórica da regressão testada. Em muitos outros campos,
entretanto, modelos teóricos são raros. Assim, os investigadores são freqüentemente forçados a explorar as variáveis independentes para que possam realizar estudos sobre a variável dependente. Obviamente, tais conjuntos de variáveis independentes são grandes. Algumas das variáveis independentes podem ser removidas seletivamente. Uma variável independente pode não ser fundamental ao problema; pode estar sujeita a grandes erros de medidas; e pode efetivamente duplicar outra variável independente da lista. Outras variáveis independentes, que não podem ser medidas, podem ser excluídas ou substituídas por variáveis que estão altamente correlacionadas com estas.

Normalmente, após uma seleção inicial, o número de variáveis independentes que permanece ainda é grande. E assim, muitas destas variáveis estarão altamente intercorrelacionadas. Portanto, o investigador geralmente desejará reduzir o número de variáveis independentes a serem usadas no modelo final. Existem várias razões para isto.

Um modelo de regressão com um número grande de variáveis independentes é caro para se utilizar. Desta forma, modelos de regressão com um número limitado de variáveis independentes são fáceis para se avaliar e estudar. Finalmente, a presença de muitas variáveis independentes altamente intercorrelacionadas, pode adicionar pouco ao poder de predição do modelo, enquanto retira suas habilidades descritivas e aumenta os erros de predição.

O problema, então, é como reduzir a lista de variáveis independentes de forma a obter a melhor seleção de variáveis independentes. este conjunto precisa ser suficientemente pequeno para que a manutenção dos custos de atualização do modelo sejam manuseáveis e a análise facilitada, e ainda, deve ser grande o suficiente de forma que seja possível uma descrição, um controle e uma predição adequados.

Os procedimentos de procura para se encontrar o melhor conjunto de variáveis independentes que deve ser empregado após o investigador ter estabelecido à forma funcional da relação de regressão, ou seja, se as variáveis dadas estão na forma linear,quadrática, etc.; se as variáveis independentes são primeiramente transformadas, como por exemplo por transformação
logarítmica; e se algum termo de interação foi incluído. Neste ponto, os procedimentos de procura são empregados para reduzir o número de variáveis independentes.

Existem muitos procedimentos de seleção, mas nenhum deles pode, comprovadamente, produzir o melhor conjunto de variáveis independentes. Não existe um conjunto ótimo de variáveis independentes, pois o processo de seleção das variáveis possui julgamentos subjetivos. Dentre os procedimentos, pode-se citar como os mais comumente usados: todas as regressões possíveis, backward, forward e stepwise. Neste trabalho vamos utilizar apenas o procedimento stepwise.

Stepwise (passo a passo): é, provavelmente, o mais amplamente usado dos métodos de pesquisa que não requerem a computação de todas as regressões possíveis. Ele foi desenvolvido para economizar esforços computacionais, quando comparado com a abordagem de todas as regressões possíveis, enquanto atinge um conjunto de variáveis independentes razoavelmente bom.
Essencialmente, este método de pesquisa computa uma seqüência de equações de regressão, adicionando ou excluindo uma variável independente em cada passo. a rotina de regressão stepwise permite que uma variável independente, trazida para dentro do modelo em um

estágio anterior, seja removida subseqüentemente se ela não ajudar na conjunção com variáveis adicionadas nos últimos estágios. Esta rotina empregada conduz a um teste para rastrear alguma variável independente que seja altamente correlacionada com variáveis independentes já incluídas no modelo. A limitação da procura da regressão stepwise é que ela presume a existência de um único conjunto ótimo de variáveis independentes e busca identificá-lo. Como notado anteriormente, não existe freqüentemente um único conjunto ótimo. Outra limitação da rotina de regressão stepwise, é que ela algumas vezes surge com um conjunto de variáveis independentes razoavelmente fraco para predições, quando as variáveis independentes estão altamente correlacionadas (Draper e Smith, 1981, p.307-312).

Após a obtenção de todas as regressões, deve-se utilizar os critérios para comparação dos modelos ajustados. Alguns critérios que podem ser usados são o R^2 (coeficiente de explicação), AIC (critério de informação de Akaike) MSE (quadrado médio dos resíduos) e C_p (estatística de Mallows).

4 RESULTADOS

4.1 Modelagem dos dados com modelo de Regressão Linear Múltiplo

Com o objetivo de determinar um modelo de regressão linear múltiplo adequado para prever o saldo gols, vamos iniciar utilizando todas as variáveis do estudo. Em seguida serão formulados outros modelos a fim de escolher as melhores variáveis para validação do modelo e também minimizar a complexidade matemática. No entanto, serão necessários a construção do modelo, análise de variância e depois a verificação dos pressupostos.

Para a escolha do modelo de regressão foi utilizada a denominação abaixo, acrescida da denominação da tabela 1:

- razão de forças (força do time A/força do time adversário) – forcA/forcB
- formação do time A quando utilizou 352 (45 partidas) – form2A
- formação do time B quando utilizou 442 (42 partidas) – form2B

4.1.1 Modelo incluindo as variáveis originais

As variáveis utilizadas e os respectivos valores dos coeficientes estão apresentados na tabela 7. As variáveis formA e formB foram substituídas por form2A e form2B, respectivamente, devido a essas terem uma melhor representação para a formação tática do time. O modelo inicial será dado por:

saldo= forcA+forcB+expeA+expeB+form2A+form2B+campo

Parâmetro	Estimativa	Erro Padrão	Estatística T	p-valor
Intercepto	-1,868	2,927	-0,638	0,526
forcA	0,546	0,108	5,040	5,18e-6
forcB	-0,469	0,056	-8,382	1,83e-11
form2A	0,707	0,723	0,978	0,332
form2B	0,615	0,732	0,840	0,404
expeA	-0,199	0,723	-0,279	0,784
expeB	0,141	0,466	0,302	0,764
campo	-0,864	0,649	-1,331	0,188

Tabela 7 – Ajuste do modelo de regressão múltipla com inclusão de todas as variáveis

Porém, precisamos fazer uma seleção de variáveis para escolher as melhores variáveis que explicarão o modelo. Para isso foi feito a escolha de variáveis pelo método stepwise. Após este método o modelo escolhido ficou da seguinte forma:

saldo = -2,090 + 0,580forcA – 0,486forcB

Parâmetro	Estimativa	Erro Padrão	Estatística T	p-valor
Intercepto	-2,090	2,336	-0,899	0,372
forcA	0,580	0,094	6,160	6,27e-8***
forcB	-0,480	0,050	-9,441	1,53e-13***

Tabela 8 – Ajuste do modelo de regressão múltipla após seleção de variáveis

Após o modelo selecionado (tabela 8), temos que verificar os resíduos a fim de testar os pressupostos de normalidade, homocedasticidade, independência e outliers a fim de validar os testes de hipóteses realizados. Esta validação pode ser feita através da figura 11.

Podemos observar na figura 11 o comportamento dos resíduos x valores preditos que mostra uma dispersão entre os pontos, este comportamento valida o pressuposto de independência dos resíduos. O segundo gráfico nos mostra que os resíduos seguem um comportamento normal. O último gráfico nos mostra os possíveis outliers, podemos ver que temos os valores 9, 21 e 36 como candidatos, mas por enquanto vamos manter essas observações no modelo.

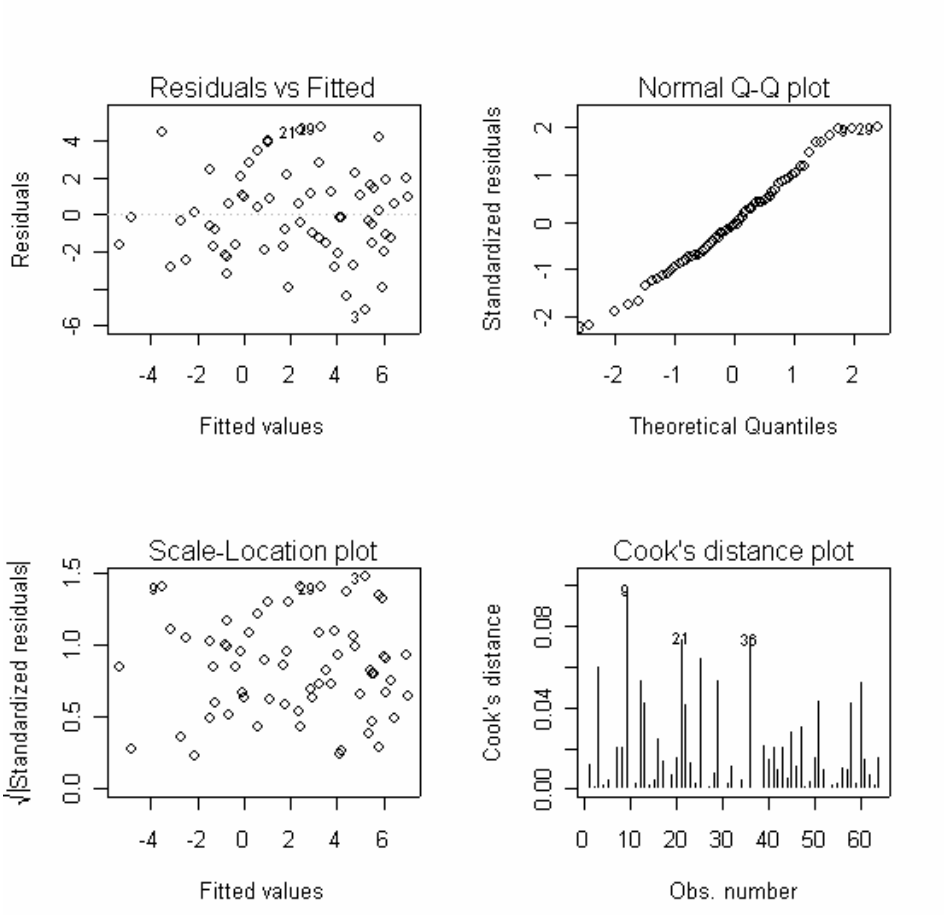


Figura 11 – Análise dos Pressupostos para o Modelo: saldo = -2,090 + 0,580forcA – 0,486forcB

4.1.2 Modelo utilizando razão de forças (forcA/forcB)

As variáveis utilizadas e os respectivos valores dos coeficientes estão apresentados na tabela 9. Neste modelo as variáveis forcA e forcB foram substituídas pela razão entre elas, ou seja, forcA/ forcB e esta transformação foi chamada de **razão**. O modelo inicial é dado por:

Saldo= razão+expeA+expeB+form2A+form2B+campo

Parâmetro	Estimativa	Erro Padrão	Estatística T	p-valor
Intercepto	-9,44916	2,491	-0,379	3,62e-4
razão forcA/forcB	7,707	0,883	8,725	4,35e-12
form2A	1,260	0,722	1,744	0,086
form2B	0,524	0,738	0,710	0,480
expeA	0,371	0,670	0,561	0,576
expeB	0,0365	0,471	0,078	0,938
campo	-0,870	0,684	-1,309	0,195

Tabela 9 – Ajuste do modelo de regressão múltipla com utilização da razão de forças (forcA/forcB)

Para a composição do modelo final, precisamos fazer uma seleção de variáveis para escolher as melhores variáveis que explicarão o modelo. Novamente foi utilizado o método stepwise. Após a aplicação do método o modelo escolhido ficou da seguinte forma:

saldo = -8,332 + 7,945razão + 1,193form2A

Parâmetro	Estimativa	Erro Padrão	Estatística T	p-valor
Intercepto	-8,332	1,244	-6,696	8,23e-9***
razão forcA/forcB	7,945	0,811	9,760	4,75e-14***
formA	1,193	0,679	1,575	0,084*

Tabela 10 – Ajuste do modelo de regressão múltipla após seleção de variáveis

Após o modelo selecionado (tabela 10), temos que verificar os resíduos a fim de testar os pressupostos de normalidade, homocedasticidade, independência e outliers. Esta análise pode ser feita através da figura 12.

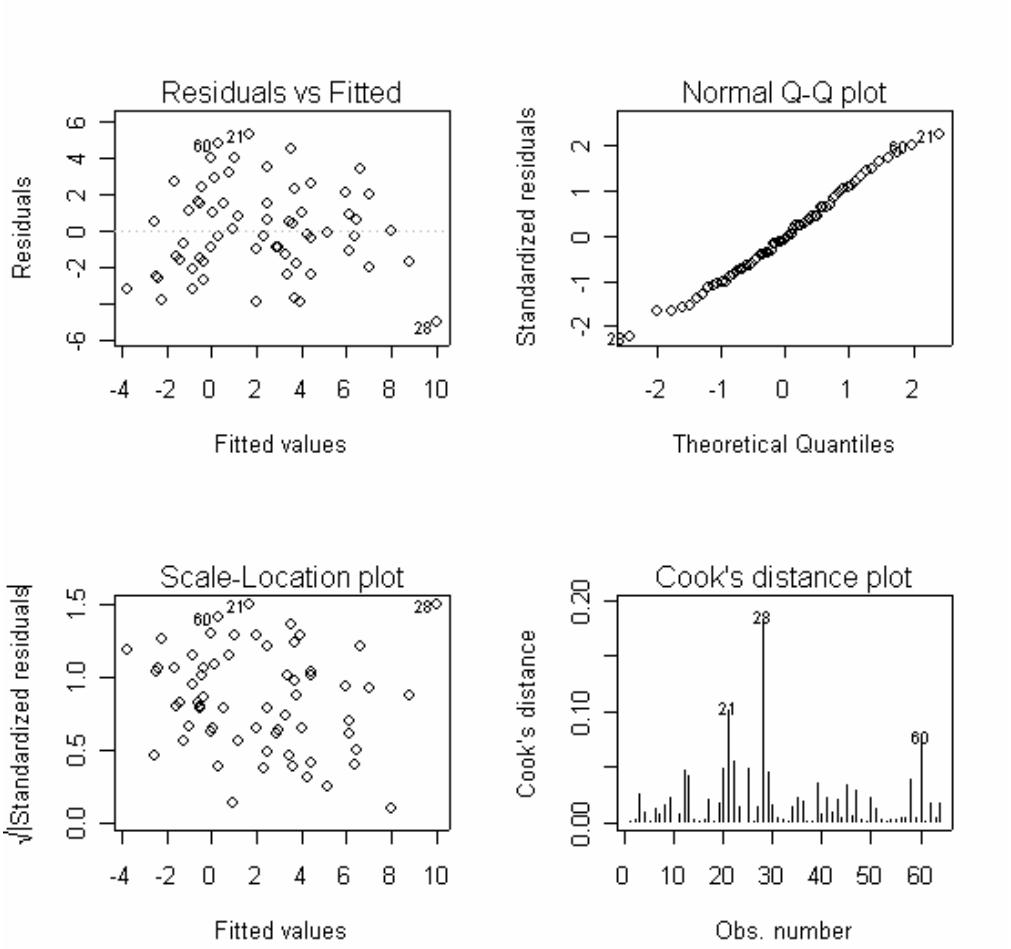


Figura 12 – Análise dos Pressupostos para o Modelo: saldo = -8,332 + 7,945razão + 1,193form2A

4.2 Comparação entre os modelos de regressão ajustados

Para a comparação entre os modelos de regressão ajustados vamos utilizar dois critérios, o AIC (Critério de informação de Akaike) e o R^2 (coeficiente de Explicação do modelo).

O AIC é uma estimativa da Logverossimilhança Negativa dos Modelos ponderada para o número de parâmetros estimados, consequentemente, o modelo com menor valor de AIC é o mais apropriado.

Sendo que:

$$AIC = -2\log(MV) + 2n_{par}$$

MV=Máxima verossimilhança

npar= número de parâmetros do modelo

Quando o ajuste é bom o modelo explica boa parte da variação total e conseqüentemente o valor de R^2 é próximo de 1, sendo que:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad 0 < R^2 < 1$$

4.2.2 Modelos comparados

Vamos comparar os dois modelos que foram ajustados em 4.1.1 e 4.2.2 identificados aqui por 1 e 2.

(1) Y = -2,090 + 0,580X1 – 0,486X2

(2) Y = -8,332 + 7,945X3 + 1,193X9 – 0,895X8

A tabela 11 mostra a comparação entre o modelo 1 e 2. Podemos observar que os maiores valores de R^2 e os menores valores de R^2 indicam que o modelo 1 é o indicado para ser o modelo final para prever o saldo de gols.

	Critério	
Model	AIC	R^2
1	117,18	0,6273
2	119,94	0,6229

Tabela 11 – Comparação entre os modelos (1 e 2) utilizando R^2 e AIC

4.3 Prevendo resultados

A tabela 12 mostra a aplicação do modelo na 17ª partida para os quatro clubes em estudo. Vale ressaltar que a modelagem foi feita a partir das dezesseis partidas durante uma temporada do campeonato disputado por esses clubes.

Clube	Modelo	Saldo Previsto	Saldo Observado
Chilics Lucidus	1	2,86	2
	2	3,12	2
Sónaveia	1	-0,86	-1
	2	-0,23	-1
Atlétic Stiletto	1	-2,51	-3
	2	-2,12	-3
Guerreiros de Seth	1	-0,92	-1
	2	0,57	-1

Tabela 12 – Previsões Estimadas para partidas realizadas após o trabalho

Podemos ver na tabela que os valores previstos pelo modelo 1 foram melhores que o modelo 2, reafirmando que o modelo 1 é melhor.

Agora vamos utilizar o melhor modelo para prever os resultados da 17ª, 18ª, 19ª e 20ª partida para os clubes.

	Partida 17		Partida 18		Partida 19		Partida 20	
	pre	obs	pre	obs	pre	obs	pre	obs
Chilics Lucidus	2,86	2	3,94	3	-2,36	-2	1,56	2
Sónaveia	-0,86	-1	1,23	2	-1,25	-2	2,38	2
Atlétic Stiletto	-2,51	-3	-2,35	2	-4,56	-3	-1,71	-1
Guerreiros de Seth	-0,92	-1	-0,28	-1	1,12	1	2,43	2

Tabela 13 – Resultados após as partidas 17, 18, 19 e 20 comparado ao modelo 1

Para facilitar a leitura foram plotados os comportamentos para cada um dos clubes

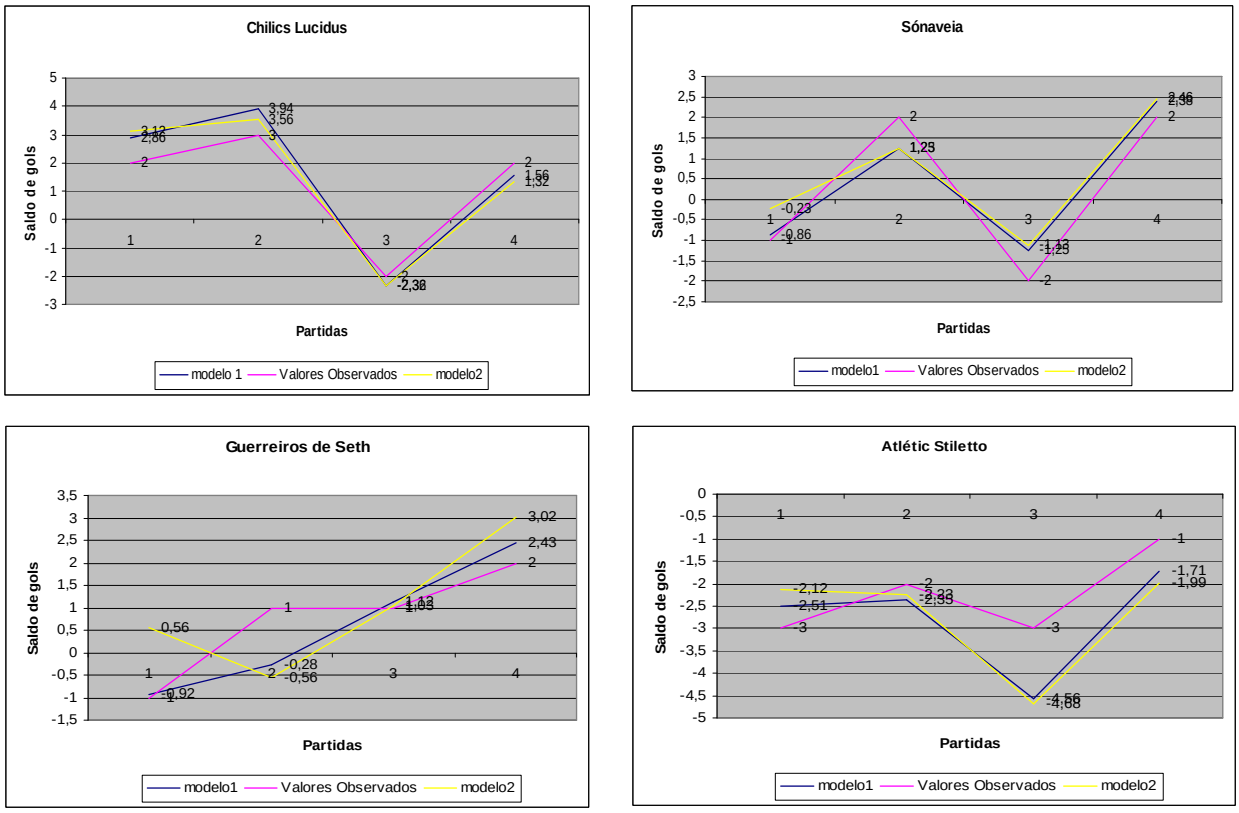


Figura 13 – Comportamento dos modelos (1 e 2) comparados com o valor observado

Podemos observar na figura 13 os valores observados (linha rosa), valores previstos pelo modelo 1 (linha azul) e valores previstos pelo modelo 2 (linha amarela). Existe um comportamento muito parecido entre os dois modelos de previsão, mas devemos optar pelo

modelo 1 visto que na maioria das vezes este modelo é mais aproximado ao valor observado.

5 CONCLUSÃO

O saldo de gols em jogos de futebol on-line pode ser estimado por meio de uma regressão linear múltipla, desde que se disponha de uma base de dados formada por uma amostra aleatória com informações do saldo de gols e as principais características observadas durante as partidas que pertençam ao modelo preditivo.

A Analise descritiva dos dados mostrou que o mando de campo não interfere no resultado das partidas mesmo quando jogada em casa. Outro fator importante observado na análise descritiva foi a forte prevalência nas formações táticas 352 e 442, mostrando que as demais formações além de pouco utilizadas não surtem efeitos importantes durante os resultados.

Os resultados obtidos com o modelo de regressão linear múltipla, que melhor representa os resultados em jogos de futebol on-line, foi eficaz e mostrou ótimos valores para o teste realizado. O fato dos dados serem provenientes de uma distribuição normal facilitou a modelagem evitando grandes transformações ou complexidades no modelo.

Por fim, sugiro como uma proposta futura a utilização de recursos mais complexos e que garantam uma maior eficiência como séries temporais, modelos lineares generalizados e aplicações de algoritmos que resolvam o problema.

Assim, o presente trabalho procurou oferecer uma contribuição para usuários dos jogos de futebol on-line.

6 REFERÊNCIAS BIBLIOGRÁFICAS

CAMPOS, H. **Estatística Experimental Não – Paramétrica**. ESALQ. 1983.

DANIEL, C.; WOOD, T. E. **Fitting equations to data.** New York: John Wiley & Sons, Inc, 1971.

DANTAS, R. A. **Engenharia de Avaliações: Introdução à metodologia cientifica.** [S.I.] São Paulo: Pini, 1998.

DRAPER, N. R. & SMIYH, H. **Applied regression analysis.** New York: John Wiley & Sons, Inc, 1981.

ELIAN, S. N. **Análise de Regressão.** São Paulo: IME, 1998.

JOHNSTON, J. **Métodos Econométricos.** São Paulo: Atlas, 1986.

Li L. M., Sander JW. National demonstration project on epilepsy in Brazil. **Arq Neuropsiquiatr.** 61(1):153–6, 2003.

NETER, J.; WASSERMAN, W. **Applied linear statistical models.** Richard D. Irwin, Inc, Ilinois, 1974.

PENNA, L. **Dicionário popular do futebol.** Rio de Janeiro: Editora Nova Fronteira, 1998.

PEREIRA, R. S. **Estatística e suas aplicações.** São Paulo: Grafosul, 1970.

R Development Core Team. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2005.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods.** 6. ed., Iowa: Ames, 1972.

7 ANEXOS

7.1 Comandos do R utilizados na análise descritiva

```
#Abrindo o conjunto de dados no R
> dados<- read.table("C:/Documents and Settings/cleibson/Desktop/dados.txt", header=T)

#listando os dados
> dados

#Listando mínimo, 1quartil, media, 3quartil, máximo
summary(dados)

#Desindexando as variáveis
> attach(dados)pl

#Dividindo a tela para plotar 4 gráficos
> par(mfrow=c(2,2))

#Plotando o saldo de gols
> hist(saldo, br=8, ylab="Número de Repetições")

#Plotando Relaçao entre saldo e forças
> par(mfrow=c(1,2))
> plot(saldo~forcA, main="Relação entre saldo e forcA")
> plot(saldo~forcB, main="Relação entre saldo e forcB")

#Boxplot relação saldo e formação
> par(mfrow=c(1,2))
> boxplot(saldo~formA, ylab="saldo", xlab="formação tática", main="timeA", ylim=c(10,-
10))
> boxplot(saldo~formB, ylab="saldo", xlab="formação tática", main="timeB", ylim=c(10,-
10))

#Boxplot relação saldo e experiência
> par(mfrow=c(1,2))
> boxplot(saldo~expeA, ylab="saldo", xlab="Nível de Experiência dos Jogadores",
main="Boxplot saldo x ExpeA")
> boxplot(saldo~expeB, ylab="saldo", xlab="Nível de Experiência dos Jogadores",
main="Boxplot saldo x ExpeB")

#Boxplot relação saldo e campo
> boxplot(saldo~campo, ylab="saldo", xlab="Mando de Campo", main="Boxplot saldo x
campo")

#Grafico de Pizzas para formação
> tformA<-table(formA)
> tformB<-table(formB)
> par(mfrow=c(1,2))
> pie(tformA, col=c(7,5,4,6), main="Formação timeA")
> pie(tformB, col=c(7,5,2,6), main="Formação timeB")
```

#Gráfico de pizza para Experiência
> texpeA<-table(expeA)
> texpeB<-table(expeB)
> par(mfrow=c(1,2))
> pie(texpeA, col=c(7,5,4,6), main="Experiência timeA")
> pie(texpeB, col=c(3,7,5,6), main="Experiência timeB")

7.2 Conjunto de dados

Clube	saldo	forcA	formA	expeA	forcB	formB	expeB	campo	formA2	formB2
a	4	32.5	b	3	23.5	a	3	c	1	1
a	5	31.5	b	3	21.5	a	3	f	1	1
a	0	31	b	3	23.5	a	3	c	1	1
a	6	28	b	2	19	a	3	f	1	1
a	4	29.5	b	3	21	a	3	c	1	1
a	-3	27	b	3	34	b	2	c	1	0
a	1	26.5	b	3	30.5	a	3	f	1	1
a	1	27	b	3	20	b	2	c	1	0
a	1	26	b	2	35	b	2	c	1	0
a	1	23	b	3	24	a	3	c	1	1
a	2	23	c	3	19.5	a	3	f	0	1
a	0	23.5	c	2	15	a	2	c	0	1
a	-2	23.5	b	3	19.5	a	2	f	1	1
a	1	22.5	b	3	20.5	a	2	c	1	1
a	5	20	b	2	12	a	2	f	1	1
a	7	23	b	3	13.5	b	2	c	1	0
b	2	26.5	c	3	28	a	3	c	0	1
b	-2	23	e	2	25.5	b	3	f	0	0
b	-3	25	b	3	29	b	3	c	1	0
b	-7	20.5	e	2	30.5	a	2	f	0	1
b	7	28.5	b	3	23	b	2	f	1	0
b	4	28	a	3	28	a	2	c	0	1
b	8	27	a	2	14	a	3	f	0	1
b	8	25	b	2	13	a	3	c	1	1
b	10	24	b	2	13	a	4	f	1	1
b	6	24.5	b	2	13.5	a	4	f	1	1
b	7	25	b	3	12	a	3	f	1	1
b	5	23	a	3	10	a	2	c	0	1
b	8	21	e	2	13	a	3	f	0	1
b	5	21.5	e	2	11	a	3	c	0	1
b	4	21.5	e	2	14.5	a	3	f	0	1
b	7	22.5	e	2	12	a	4	c	0	1
c	2	27.5	b	2	22.5	b	3	f	1	0
c	-2	21	b	3	24.5	b	4	c	1	0
c	-5	22.5	b	2	33.5	a	3	f	1	1
c	2	28.5	e	2	17.5	a	2	c	0	1
c	-2	23	b	2	25.5	b	3	f	1	0
c	4	25	b	2	17	a	2	f	1	1
c	6	25	b	2	19.5	b	2	c	1	0
c	2	25.5	b	2	18.5	a	3	c	1	1
c	3	22.5	b	2	23	b	2	c	1	0
c	-2	24	b	2	26	a	4	f	1	1

c	-4	22.5	b	2	23.5	b	3	f	1	0
c	1	23	b	2	26	a	4	c	1	1
c	-6	19	b	2	24.5	b	3	f	1	0
c	-3	21.5	a	2	22	a	3	f	0	1
c	-5	19	a	2	23.5	b	3	c	0	0
c	0	18	a	2	17.5	a	2	f	0	1
d	2	29.5	b	3	24.5	b	3	c	1	0
d	4	30	b	2	28.5	b	3	c	1	0
d	-3	30	e	2	33	b	4	c	0	0
d	-1	28.5	b	2	27	b	3	f	1	0
d	4	27	b	3	21	a	3	c	1	1
d	2	22	b	2	16	a	2	f	1	1
d	3	20	b	2	16.5	a	1	c	1	1
d	2	25	b	2	19	a	2	c	1	1
d	2	28	b	3	20	b	2	f	1	0
d	5	29	b	3	30	a	3	c	1	1
d	5	28	b	3	16	a	2	f	1	1
d	5	28	e	3	27	a	2	c	0	1
d	0	29.5	e	3	25.5	a	2	f	0	1
d	7	27	b	2	18.5	a	2	c	1	1
d	1	26.5	b	2	26.5	b	3	f	1	0
d	9	29	b	3	16	f	3			

Tabela 14 – Conjunto de dados utilizados no trabalho