VIII ERMAC-R3

8º Encontro Regional de Matemática Aplicada e Computacional

20 a 22-Novembro-2008

Universidade Federal do Rio Grande do Norte - Natal/RN

Análise da invencibilidade dos clubes no "Brasileirão 2008" via Modelos Lineares Generalizados

Kleber Napoleão Nunes de Oliveira Barros,

Depto. de Biometria e Estatística Aplicada, UFRPE,

17033-360, Recife, PE

E-mail: napoleaok@hotmail.com,

Patrícia Silva Nascimento

Depto. de Estatística, UFPE,

15054-400, Recife, PE

E-mail: patriciajp1408@hotmail.com.

Resumo

Muito se diz a respeito da imprevisibilidade no futebol, sobretudo em relação ao campeonato brasileiro que apresenta uma grande variabilidade. Mas, até que ponto o futebol é realmente imprevísivel? Variações econômicas, políticas, qualidade técnica da equipe, jogar dentro ou fora de "casa", pontuação na tabela, ter ganho o jogo anterior, atuar em outro campeonato, dentre outros fatores, determinam a probabilidade de um time vencer. Na literatura, alguns trabalhos em economia ligados ao esporte tentam explicar teorias econômicas através do esporte. Em estatística aparecem trabalhos ligados a análise descritiva e teoria da regressão. Este artigo objetiva introduzir Modelos Lineares Generalizados (MLGs) para modelar a invencibilidade. O MLG é definido através da família de distribuições exponenciais para a variável resposta, um conjunto de variáveis explicativas (componente sistemático) e uma função de ligação adequada. Obteve-se as variáveis explicativas no site da Confederação Brasileira de Futebol (CBF) e Probabilidades no Futebol da UFMG e referem-se a situação de cada time no "brasileirão 2008 série A" até a 27° rodada. Analisou-se vários modelos da família exponencial, concluindo que o modelo Gama com função de ligação 'identidade' se adequou melhor através do critério de informação de Akaike e análise de resíduos.

Palavras-chave

Modelos Lineares Genearalizados, Futebol, Invencibilidade, Critério de Informação de Akaike.

Introdução

É muito comum, torcedores, técnicos, jogadores e pessoas direta ou indiretamente ligadas ao futebol, se perguntarem sobre as chances de um time ganhar uma partida, ganhar um campeonato, se classificar para uma outra competição ou até mesmo de ser rebaixado. Muito se comenta sobre estatística de partidas e torneios, porém poucas são as análises que realmente levam considerações mais elaboradas, tentando obter métodos adequados para fazer previsões estatísticas dos resultados. Na literatura encontramos trabalhos ligados a econometria que tentam explicar a influência da política e das finanças do estado no qual o clube está inserido e do próprio clube no seu desempenho ao longo das partidas [1].

O Campeonato Brasileiro de Futebol, em seus moldes atuais, dispõem de dois turnos com 19 rodadas e 10 jogos cada. Desta forma, o campeonato tem ao todo 380 jogos. Por ser de pontos corridos, este se torna um excelente problema para se fazer inferências sobre quaisquer variáveis mensionadas acima. Contudo, a quantidade de fatores envolvidos em um campeonato com o tamanho do "Brasileirão" torna qualquer análise incompleta e imprecisa. Tanto que, se quiséssemos apenas guardar as combinações de vitórias, empates e derrotas entre os times teriamos o número 3^{380}, algo muito superior a quantidade de átomos do do nosso sistema solar [10]. Somado ao contexto da partida, fatores extra-campo como variações econômicas, políticas, ter ganho o jogo anterior, atuar em outro campeonato, dentre outros, também interagem na probabilidade de um time ser vitorioso. Segundo Pollard, Silva & Medeiros (2008) a distância percorrida pelo time visitante tem efeito pequeno, mas significante no resultado do jogo ($p < 0,01$) correspondendo a 0,115 de gol a favor do time da casa para cada 1000km viajados pelo visitante [8]. Tudo isso torna o futebol imprevísivel, ou como se diz: uma "caixinha de surpresas". Mas, até que ponto o futebol é realmente imprevísivel? Utilizando-se técnicas estatísticas robustas, podemos chegar a modelos parcimoniosos, isto é, com alto grau de precisão e uso de poucas variáveis (baixo custo computacional). Pereira et. al (2004) apresenta como resultado um grau de correlação significativo entre receitas, custos, despesas e resultados do clube com a sua posição no campeonato, sugerindo que o desempenho financeiro do clube está diretamente relacionado à sua performance esportiva [9].

Para este artigo designamos a variável resposta INVENC, definida como a maior sequência de partidas em que um time ganha ou empata. O objetivo deste artigo, então, é modelar a invencibilidade dos times no "Brasileirão 2008" utilizando Mode-

los Lineares Generalizados. Obteve-se as variáveis explicativas no site da Confederação Brasileira de Futebol (CBF) [2] e Probabilidades no Futebol da UFMG [10] que referem-se a situação de cada time no "brasileirão 2008 série A" até a 27° rodada. As variáveis utilizadas foram as seguintes (Tabela 1):

Tabela 1: Descrição das variáveis em estudo

Variáveis	Descrição
y	Maior sequência de invencibilidade
x_1	Posição na tabela (POS)
x_2	Pontos ganhos (PON)
x_3	Número de vitórias (VIT)
x_4	Número de derrotas (DER)
x_5	Número de empates (EMP)
x_6	Gols marcados (GOLM)
x_7	Gols sofridos (GOLS)
x_8	Saldo de gols (SALG)
x_9	Aproveitamento (APR)
x_{10}	Total de público (TPU)
x_{11}	Média de público (MPU)
x_{12}	Total de arrecadação (TAR)
x_{13}	Média de arrecadação (MAR)
x_{14}	Prob. de vitória c/ mandante (PVM)
x_{15}	Prob. de empate c/ mandante (PEM)
x_{16}	Prob. de derrota c/ mandante (PDM)

Metodologia

Os Modelos Lineares Generalizados (MLGs) foram primeiramente proposta por Nelder e Wedderburn em 1972, o qual combina, dentro de uma mesma estrutura, diversos modelos usuais, como modelos de regressão linear, logística, modelos log-lineares, análise de variância e covariância, entre outros [6]. Os MLGs evitam a aplicação de transformações para alcançar normalidade, linearidade e homogeinidade da variância, como exemplos de dados que podem ser analizados temos, dados de contagem ou dicotômicos.

Seja Y uma variável resposta associada a um conjunto de variáveis explicativas $x_1, ..., x_p$. Para uma amostra de n observações (y_i, x_i), o modelo envolve três componentes dados por [4]:

- Componente aleatório: Seja um vetor de observações $y = (y_1, ..., y_n)^T$ referente às realizações das variáveis aleatórias $Y = (Y_1, ..., Y_n)^T$, independentes e identicamente distribuídas (iid), com médias $\mu = (\mu_1, ..., \mu_n)^T$. Y segue uma distribuição da família exponencial com médias $E(Y_i) = \mu_i$, com fdp de Y dada por:

$$f(y|\theta, \phi) = exp\left\{\phi^{-1}[y\theta - b(\theta)] + c(y, \phi)\right\}, \quad (1)$$

em que θ_i é o parâmetro canônico e ϕ_i é parâmetro de dispersão.

- Componente sistemático (variáveis explicativas): as variáveis explicativas com estrutura linear entram no modelo da seguinte forma:

$$\eta = X\beta, \quad (2)$$

em que $X = (x_1, ..., x_n)^T$ é a matriz do modelo conhecida de posto p, $\beta = (\beta_1, ..., \beta_p)^T$ é o vetor de parâmetro e $\eta = (\eta_1, ..., \eta_n)^T$ é o preditor linear.

- Função de ligação: função que relaciona o componente aleatório ao sistemático, isto é,

$$\eta_i = g(\mu_i) \quad (3)$$

em que $g(\cdot)$ é uma função monótona e diferenciável.

Utilizou-se o método da máxima verosimilhança, para obter os estimadores dos parâmetros do modelo, através do método iterativo Escore de Fisher:

$$\beta^{(m+1)} = \beta^{(m)} + U^{(m)}(K^{(m)})^{-1} \quad (4)$$

em que $\beta^{(m)}$ são as estimativas dos parâmetros, $U^{(m)}$ é o vetor de primeiras derivadas da função de log-verossimilhanças e $(K^{(m)})^{-1}$ é a inversa da matriz de informação, todos avaliados no passo m. A matriz de informação de Fisher é dada por $\kappa_{r,s} = -E\left[\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right]$.

Este método consiste em determinar o máximo absoluto do logaritmo da função verosimilhança.

A análise de diagnóstico compreende a detecção de discrepâncias isoladas (pontos atípicos) que podem ser caracterizadas por serem inconsistentes e/ou serem influentes [3].

- Para detectar os possíveis pontos de Alavanca, utiliza-se os elementos da diagonal principal da matriz $H = \hat{W}^{1/2}X(X^T\hat{W}X)^{-1}X^T\hat{W}^{1/2}$, h_{ii}, também conhecida como matriz de projeções ou matriz *hat*, em que $\hat{W}^{1/2}z$ é o vetor resposta. Um ponto de alavanca possui um perfil diferente dos demais em relação aos valores das variáveis explicativas [7]. Espera-se que observações distantes do espaço formado pelas variáveis explicativas apresentem valores apreciáveis de h_{ii}. Se $h_{ii} > 2p/n$ temos que y_i é ponto de alavanca.

- No caso de ponto influente (aquele que exerce um pesso desproporcional nas estimativas dos parâmetros do modelo), utiliza-se a Estatística Modificada de Cook (T_i). Se $T_i > 2(p/n)^{1/2}$, y_i será influente.

- Ponto aberrante é aquele que apresenta perfil diferente das demais observações em relação aos valores da variável resposta. Para detectarmos pontos aberrantes, graficamente, utiliza-se o resíduo padronizado versus o índice das observações.
- O Gráfico de Probabilidade Meio-Normal (qq-plot) com Envelope Simulado é construído baseado nos resíduos padronizados, a ocorrência de pontos próximos ou fora dos envelopes indicam que o modelo não está apropriado.

O Critério de Informação de Akaike (AIC) é uma medida utilizada para selecionar o modelo mais parcimonioso entre os ajustados, ou seja, que esteja bem ajustado e com um número reduzido de parâmetros [5]. O AIC é dado por:

$$AIC = -2l(\hat{\beta}) + 2p \qquad (5)$$

em que $l(\hat{\beta})$ é o logaritmo da função de verossimilhança dos dados e p é o número de parâmetros. O modelo escolhido será o que tiver menor AIC.

Utilizou-se o gráfico dos resíduos versus o preditor liner η para estudar a adequação da função de variância. Para que a função de variância seja adequada, este deve apresentar pontos dispersos ou simplesmente sem aparente tendência (linear, quadrática, logarítmica, entre outras). O teste da adequação da função de ligação consiste em se adicionar $\hat{\eta} \otimes \hat{\eta}$ como uma variável explanatória extra e examinar a mudança ocorrida no desvio, o que equivale ao teste da razão de verossimilhanças [3]. Se ocorrer uma diminuição, há evidência de que a função de ligação é insatisfatória e que se deve incluir u no modelo.

Resultados

Analisou-se diversos modelos e escolheu-se o modelo com distribuição Gama e função de ligação "identidade" por obter o menor AIC (65,625). As seguintes variáveis foram excluídas : EMP, SALG (não convergiram), PON, GOLM, GOLS, PEM, PDM (não significativas). Na Tabela 2 tem-se as estimativas dos coeficientes do modelo ao nível de significância de 10%.

Na Tabela 2 verifica-se que as variáveis DER $(-2,27)$, APR $(-0,62)$, MPU $(-6,58\text{e-}03)$, TAR$(-2,66\text{e-}05)$ e PVM$(-0,36)$ influenciam negativamente para o modelo, sendo DER a variável de maior influência negativa. As variáveis que influenciam positivamente foram POS (0,62), VIT (5,88), TPU (5,27e-04) e MAR (3,13e-04), sendo VIT a mais influente positivamente.

Na Figura 1 verifica-se três pontos (observações) de influência 4 (Flamengo), 10 (Coritiba) e 12 (Atlético-MG), sendo a observação 12 a mais influente. As observações 1 (Palmeiras), 13 (Náutico) e 16 (Atlético-PR) são pontos de alavanca para o modelo. Nota-se também que a observação 2 (Grêmio) é o ponto aberrante (outlier), ou seja, apresenta perfil diferente dos demais em relação aos valores da variável resposta (invencibilidade). A Figura 2 apresenta o gráfico de diagnóstico do modelo onde confirmamos que essa observação é ponto aberrante, dessa forma a retiramos do modelo. A Figura 3 apresenta o qq-plot onde observa-se um ponto fora da banda de confiança produzida pelo envelope simulado, o que indica que o modelo deve ser revisto.

Tabela 2: Estimativa dos coeficientes para o modelo Gama com função de ligação "identidade"

Coeficiente	Estimativa	Erro padrão	P-valor
β_0	2,90e+04	9,54e+03	0,0187
β_1	0,62	0,27	0,0521
β_3	5,88	1,26	0,0023
β_4	$-2,27$	0,64	0,0092
β_9	$-0,62$	0,31	0,0909
β_{10}	5,27e-04	2,04e-04	0,0360
β_{11}	$-6,58\text{e-}03$	2,65e-03	0,0420
β_{12}	$-2,66\text{e-}05$	1,28e-05	0,0752
β_{13}	3,13e-04	1,65e-04	0,0995
β_{14}	$-0,36$	8,33e-02	0,0037

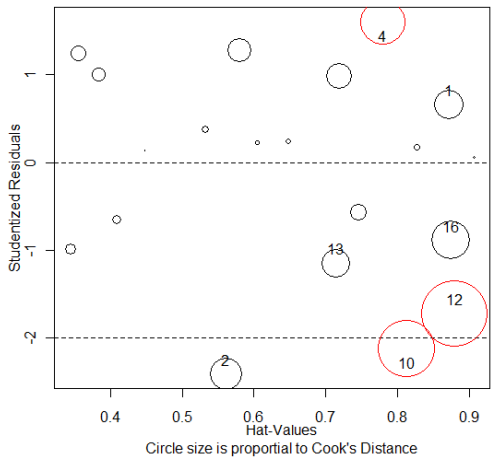

Figura 1: Gráfico de pontos de influência

Após a retirada da observação discrepante (Grêmio) para o mesmo modelo e mesma função de ligação, observou-se uma diminuição do AIC (51,362), ou seja, o modelo tornou-se mais parcimonioso que o anterior. Para verificar a adequação da função de variância utilizamos o método empírico do esboço dos resíduos versus o preditor liner η. Na Figura 4 aparecem os resíduos versus η (preditor linear). Como os pontos não parecem apresentar uma tendência (linear, quadrática, logarítmica, ...), concluimos que a função de variância é adequada

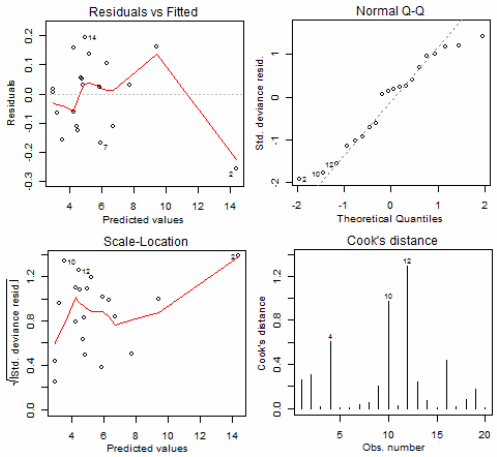

Figura 2: Gráficos de diagnóstico do modelo

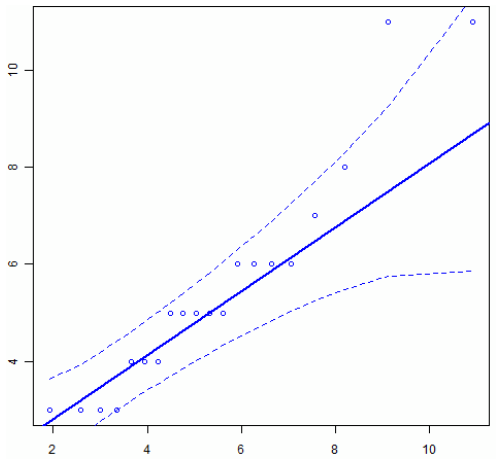

Figura 3: Qq-plot do modelo ajustado

para o modelo. No teste da adequação da função de ligação, a inclusão de η^2 não proporcionou uma diminuição acentuada no desvio residual que passou de 0,158 para 0,124, além disso a inclusão de η^2 fez com que os coeficientes deixassem de ser significativas, sugerindo que η^2 não deve permanecer no modelo.

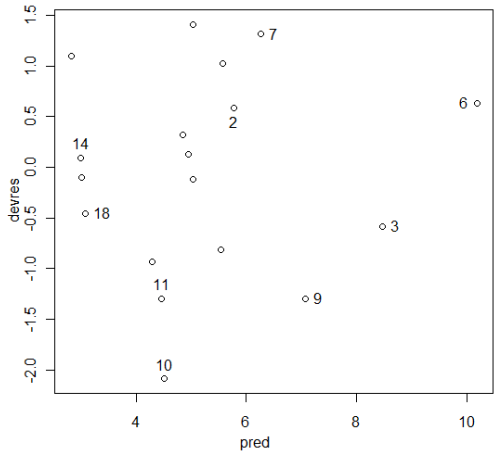

Figura 4: Gráfico de dispersão utilizado na verificação da adequação da função de variância.

Tabela 3: Estimativa dos coeficientes para o modelo Gama com função de ligação "identidade"

Coeficiente	Estimativa	Erro padrão	P-valor
β_0	4,45e+04	1,09e+04	0,0066
β_1	0,94	0,26	0,0120
β_3	7,85	1,41	0,0014
β_4	$-2,82$	0,60	0,0032
β_9	$-0,69$	0,26	0,0370
β_{10}	7,73e-04	2,03e-04	0,0090
β_{11}	$-9,72$e-03	2,63e-03	0,0102
β_{12}	$-4,02$e-05	1,23e-05	0,0170
β_{13}	4,82e-04	1,57e-04	0,0222
β_{14}	$-0,50$	9,66e-02	0,0021

Na Tabela 3 tem-se as estimativas dos coeficientes do modelo ao nível de significância de 5%. Nota-se que todas as variáveis do modelo continuam influenciando o modelo da mesma forma, porém as estimativas se afastaram, em módulo, do zero. Portanto, para cada unidade acrescida a variável VIT (vitórias) a variável resposta (sequência de invencibilidade) é incrementada 7,85 unidades. Por outro lado, a cada unidade acrescida a variável DER (derrotas) a variável resposta perde 2,82 unidades. Isto é condizente com o senso comum.

Na Figura 4 verifica-se que a observação 4 (Flamengo) deixou de ser influente após a retirada da observação 2. As obsevações 10 (Coritiba) e 12 (Atlético-MG) continuam sendo influentes.

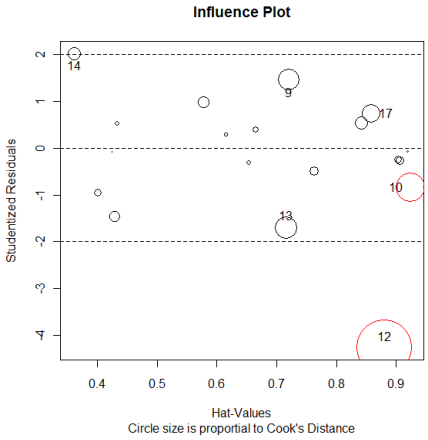

Figura 5: Gráfico de pontos de influência

A observação 13 (Náutico) continua sendo ponto de alavanca, porém 1 (Palmeiras) e 16 (Atlético-PR) foram substituídos por 9 (Internacional) e 17 (Ipatinga). Por esta análise, nota-se também que a observação 14 (Santos) é o ponto aberrante (outlier). No entanto a Figura 5 apresenta uma série de gráficos de diagnóstico em que o Santos não aparece como discrepante. Ao retirá-lo do modelo, constatamos que as estimativas não melhorarão, decidimos então continuar com o modelo ajustado. A Figura 6 apresenta o qq-plot onde observa-se que todos os pontos estão dentro da banda de confiança produzida pelo envelope simulado, confirmando a adequação do modelo. Dessa forma o modelo é dado por:

$$\hat{\mu} = 44500 + 0,94x_1 + 7,85x_3 - 2,82x_4$$
$$-0,69x_9 + 0,0008x_{10} - 0,0097x_{11}$$
$$-0,00004x_{12} + 0,0005x_{13} - 0,50x_{14}$$

em que:

Tabela 4: Descrição das variáveis do modelo final

Variáveis	Descrição
x_1	Posição na tabela (POS)
x_3	Número de vitórias (VIT)
x_4	Número de derrotas (DER)
x_9	Aproveitamento (APR)
x_{10}	Total de público (TPU)
x_{11}	Média de público (MPU)
x_{12}	Total de arrecadação (TAR)
x_{13}	Média de arrecadação (MAR)
x_{14}	Prob. de vitória c/ mandante (PVM)

Com parâmetro de dispersão $\phi = 0,027$ e 6 graus de liberdade.

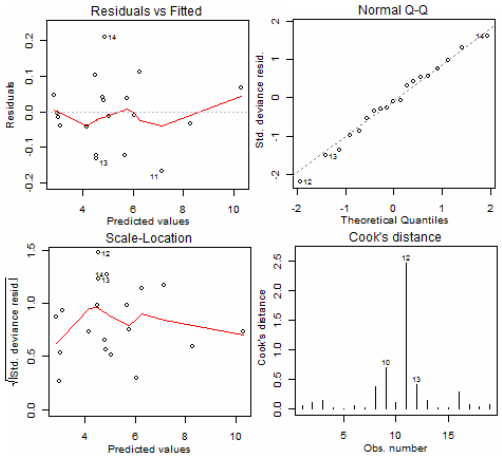

Figura 6: Gráficos de diagnóstico do modelo

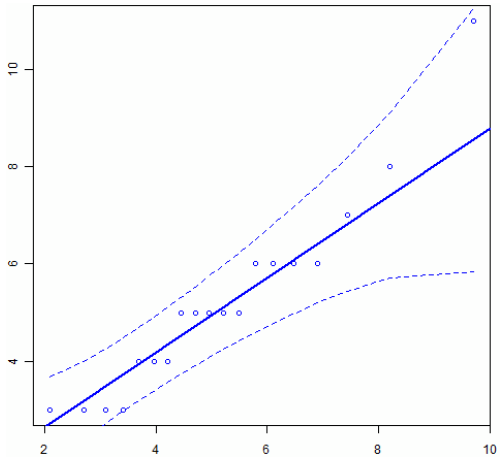

Figura 7: Qq-plot do modelo ajustado

Conclusões

Este artigo utilizou técnicas de Modelos Lineares Generalizados (MLGs) para analizar a variável resposta invencibilidade em função de variáveis explicativas obtidas nos sites da Confederação Brasileira de Futebol (CBF) e Probabilidades no Futebol da UFMG. Optou-se pelo modelo Gama com função de ligação "identidade", pois se adequou melhor aos dados com $AIC = 65{,}625$.

Com base na análise dos resíduos, observou-se uma observação aberrante (Grêmio), excluiu-se esta observação, e escolheu-se o modelo, verificou-se uma diminuição no AIC (51,362) indicando uma melhor adequação do modelo. Fez-se um teste de adequação da função de variância e função de ligação, o qual revelou que a função de ligação era adequada e os desvios estavam bem ajustados, pois ao aplicar o teste, não houve diminuição significativa dos mesmos. Dessa forma, no modelo escolhido verificou-se que o número de vitórias e a posição na tabela têm importante influência na invencibilidade de uma equipe. No modelo final o número de vitórias contribui com 7,85 unidades para a variável resposta, enquanto a posição na tabela contribui com 0,94, isto é, um time que está a uma posição acima na tabela tem 94% mais chance de continuar sem perder que seu rival. Por outro lado, o número de derrotas influencia negativamente o modelo com um decréscimo de 2,82 unidades, ou seja, se uma equipe tem uma derrota a menos que outra, está deve ter, aproximadamente, três unidades da variável resposta (invencibilidade) a mais que a segunda. Apesar da média de arrecadação e do total de público ter influência na invencibilidade de uma equipe, esta é pequena, pois nem sempre se joga em casa. O aproveitamento, a média de público, o total de arrecadação e a probabilidade de vencer como mandante tem influencia negativa, pelo mesmo motivo mencionado anteriormente.

De um modo geral constatou-se que a teoria dos Modelos Lineares Generalizados mostrou-se um método eficaz para modelar a série de vitórias e empates das equipes do Campeonato Brasileiro 2008 - Série A pelas variáveis descritas anteriormente.

Agradecimentos

A Universidade Federal Rural de Pernambuco (UFRPE) e a Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) pelo apoio financeiro.

Referências

[1] A. F. ARAUJO-JUNIOR, C. D. SHIKIDA, L. M. MONASTERIO, (2003). Abrindo a "caixinha de surpresas": uma análise econométrica do futebol brasileiro, Ibmec MG Working Paper WP3, disponível em $< http : //www.ceaee.ibmecmg.br/wp/wp3.pdf >$, acesso em 30 de setembro de 2008

[2] Confederação Brasileira de Futebol (CBF), disponível em $< http : //www.cbf.com.br/ >$, acesso em 03 de outubro de 2008.

[3] G. M. CORDEIRO, C. G. B. DEMÉTRIO, (2007). Modelos lineares generalizados. Minicurso para o 12° SEAGRO e a 52° Reunião Anual da RBRAS, UFSM, Santa Maria, RS.

[4] G. M. CORDEIRO, E. A. L. NETO, (2006). Modelos Paramétricos. UFRPE/UFPB, Recife.

[5] P. MCCULLAGH, J. A. NELDER, (1989). Generalized Linear Models. Chapman and Hall, London, 2.ed.

[6] J. A. NELDER, R. W. M. WEDDERBURN, (1972). Generalized linear models. Journal of the Royal Statis-tical Society, A, 135, 370-384.

[7] G. A. PAULA, (2004). Modelos de Regressão com Apoio Computacional. IME/USP, São Paulo.

[8] R. POLLARD, C. D. SILVA, N. C. MEDEIROS, (2008). Home advantage in football in Brazil: differences between teams and the effects of distance traveled. The Brazilian Journal of Soccer Science, Jan-Jul, 01(1): 03-10.

[9] C. A. PEREIRA, A. J. REZENDE, L. J. CORRAR, E. M. LIMA, (2004). A Gestão estratégica de clubes de futebol: Uma análise da correlação entre performance esportiva e resultado operacional, disponível em $< http : //www.congressoeac.locaweb.com.br/artigos42004/336.pdf >$, acesso em 30 de setembro de 2008.

[10] Probabilidades no Futebol da UFMG, disponível em $< http : //www.mat.ufmg.br/futebol/ >$, acesso em 03 de outubro de 2008.