Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 70 número 232 janeiro/junho 2009

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 70, n. 232, p. 1-91, jan./jun. 2009

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística - ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Francisco Louzada-Neto
Editor responsável - RBEs - IBGE.
Rua André Cavalcanti, 106 - Santa Teresa
20231-050 - Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Sandra Cavalcanti de Barros - Escola Nacional de Ciências Estatísticas - ENCE/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2009.

Capa
Renato J. Aguiar - Coordenação de *Marketing*/CDDI/IBGE

Ilustração da Capa
Marcos Balster - Coordenação de *Marketing*/CDDI/IBGE

Revista brasileira de estatística / IBGE, - v.1, n.1 (jan./mar.1940), - Rio de Janeiro : IBGE, 1940.
v.

Trimestral (1940-1986), semestral (1987-).
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).
Co-edição com a Associação Brasileira de Estatística a partir do v.58.
ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

Gerência de Biblioteca e Acervos Especiais CDU 31(05)
RJ-IBGE/88-05 (rev.2009) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor

Este primeiro volume da RBEs do ano de 2009 é composto por quatro artigos. O primeiro artigo, de autoria de Viviana Giampaoli, Maria Cristina Falcão Raposo e Silvia Teresa Freire Torres, que apresenta critérios de seleção de modelos para o modelo de regressão beta. O segundo artigo, de autoria de Cláudia Patrícia Costa de Macedo e Dione Maria Valença, considera a identificação de fatores de risco em pacientes com Câncer de mama via modelo de Cox. O terceiro artigo, de autoria de Guilherme Guimarães Moreira e Marcelo Azevedo Costa, propõe modelos de redes neurais generalizadas. O quarto artigo, de autoria de Célia Mendes Carvalho Lopes, Airlane Pereira Alencar, Franco de Sá Barroso Lippi e Flávio Hideki Yamamoto, apresenta os resultados de um estudo da produção de álcool no Brasil via técnicas de séries temporais.

Aproveito a oportunidade para esclarecer que, por solicitação da diretoria da ABE, continuarei editorando este periódico. Também, agradeço a colaboração de Ismenia Blavastsky (Editora Executiva) e todos os Editores Associados, dos revisores do periódico, autores e à equipe do IBGE.

Uma excelente leitura.

Francisco Louzada-Neto
Editor Responsável

Critérios de seleção de modelos para o modelo de regressão Beta

Viviana Giampaoli[1]
Maria Cristina Falcão Raposo[2]
Silvia Teresa Freire Torres[3]

Resumo

O modelo de regressão Beta possui grande aplicabilidade prática, em particular, na modelagem de taxas e proporções e, tal como nos demais modelos de regressão, também são requeridos métodos que determinem qual o melhor modelo. O presente trabalho tem como objetivo principal implementar e avaliar o desempenho de diferentes critérios de seleção de modelos para o modelo de regressão Beta. Para tal, mediante diferentes estudos de simulações de Monte Carlo, analisamos alguns critérios selecionados levando em consideração suas propriedades assintóticas, os quais foram obtidos por meio da função de máxima verossimilhança. Os resultados das simulações revelaram que os desempenhos dos referidos critérios dependem da especificação do modelo e também do tamanho da amostra. Apresentamos ainda uma aplicação relacionada ao Índice de Desenvolvimento Humano, que é uma variável adequada á modelagem em estudo, visto que seus valores variam no intervalo (0,1).

[1] Universidade de São Paulo
[2] Universidade Federal de Pernambuco
[3] Universidade Católica de Pernambuco

1. Introdução

Em análises estatísticas, em particular na análise de regressão, surge sempre uma pergunta importante: qual é o melhor modelo? Assim, um dos objetivos principais da seleção de modelos é atingir um equilíbrio entre uma melhora no ajuste e a complexidade do modelo.

Para decidir qual é o modelo mais apropriado dentre um conjunto de modelos candidatos, foram criados os denominados critérios de seleção de modelos. Entre os mais conhecidos, podemos citar o critério de seleção de modelos pseudo R2 (R2p), o Cp de Mallows (MALLOWS, 1973), o critério de informação de Akaike (AIC, AKAIKE, 1974) e o critério de informação bayesiano (BIC, AKAIKE, 1978) entre outros. Existem artigos mais recentes como os de Rao e Wu (2005) e Kuha (2004), os quais analisam o desempenho dos modelos utilizando funções de penalidade. Uma excelente referência deste assunto para modelos de regressão normal e de séries temporais é o livro de McQuarrie e Tsai (1998). Estes autores destacam que um critério nem sempre é melhor que outro. O fato de que certos critérios têm um desempenho melhor que outros, para um modelo específico, foi o que motivou a nossa pesquisa.

Neste trabalho, apresentamos uma análise do desempenho de vários critérios para um modelo particular proposto por Ferrari e Cribari-Neto (2004), o chamado modelo Beta, detalhado na segunda seção. Na terceira seção, definimos os diferentes critérios para os modelos Beta; na quarta, discutimos os resultados das simulações de Monte Carlo para dois modelos específicos, apresentando o desempenho de cada um deles em relação às eficiências observadas segundo a distância de L2. Na quinta seção, são apresentados os resultados de uma aplicação, usando como variável a ser explicada o Índice de Desenvolvimento Humano Municipal, em função de algumas variáveis explicativas e, finalmente, as principais conclusões são discutidas na sexta seção.[4]

[4] Neste artigo, para realizar os ajustes dos correspondentes modelos, os cálculos dos critérios, bem como as simulações, desenvolvemos programas computacionais através da linguagem matricial de programação Ox em sua versão 3.40 para plataforma computacional *Windows*, desenvolvida por Jurden A. Doornik. Ox pode ser obtida gratuitamente para uso acadêmico e está disponível em http//www.doornik.com.

2. O Modelo Beta

Para analisar a relação existente entre variáveis aleatórias, é bastante utilizado o modelo de regressão linear normal padrão, pela facilidade de se encontrar ferramentas de ajuste e pela sua interpretabilidade. No entanto, se a variável resposta estiver restrita ao intervalo (0,1), o mesmo não deve ser aplicado, visto que podemos obter valores ajustados fora deste intervalo. Uma solução para este problema é transformar a variável resposta. Entretanto, defrontamo-nos com algumas desvantagens, sendo a principal a difícil interpretação dos parâmetros. A fim de reverter estas desvantagens, Ferrari e Cribari-Neto (2004) propuseram um modelo, chamado modelo Beta, baseado na suposição que a variável resposta tem distribuição Beta.

Como na análise de regressão, é normalmente útil modelar a média da variável resposta, bem como definir o modelo de forma que contenha um parâmetro de precisão. Ferrari e Cribari-Neto (2004) propõem uma parametrização que permite obter uma estrutura de regressão associada a um parâmetro de precisão. A densidade da distribuição Beta pode ser escrita da seguinte forma

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} I_{(0,1)}(y),$$

onde $p > 0$ e $q > 0$ são parâmetros que indexam a distribuição Beta e $\Gamma(.)$ é a função gama. Sendo a função gama dada por

$$\Gamma(p) = \int_0^\infty y^{p-1} e^{-y} dy,$$

e

$$I_{(0,1)}(y) = \begin{cases} 1 \, se \; y \in (0,1) \\ 0 \, se \; y \notin (0,1). \end{cases}$$

A média e a variância da variável *y* são, respectivamente,

$$E(y) = \frac{p}{p+q}$$

e

$$\text{var}(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

Seja $\phi > 0$ fazendo $\mu = p/(p+q)$ e $\phi = p+q$, isto é, $p = \mu\phi$ e $q = (1-\mu)\phi$ temos a seguinte parametrização

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y), \quad (1)$$

logo a média e a variância de *y* são, respectivamente:

$$E(y) = \mu,$$

e

$$\text{var}(y) = \frac{V(\mu)}{1+\phi},$$

onde $V(\mu) = \mu(1-\mu)$, tal que μ é a média da resposta e ϕ pode ser interpretado como um parâmetro de precisão. Percebe-se que quanto maior for o valor de ϕ tanto menor será a variância de *y*, fixando-se μ.

Como já foi citado, trabalhamos aqui com variável resposta restrita ao intervalo (0,1). Quando a resposta está restrita ao intervalo (a; b), onde a < b são escalares desconhecidos, o modelo é também adequado e os resultados correspondentes continuam válidos. Então, ao invés de modelarmos *y*, utilizaremos $(y-a)/(b-a)$ que ficará, portanto, definida no intervalo (0; 1): O modelo proposto por Ferrari e Cribari-Neto (2004) é descrito a seguir.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, as quais seguem densidades em (1) com média μ_t, $t = 1, \ldots, n$ e precisão ϕ desconhecidas. Assuma

que a média de y_t no modelo pode ser escrita como

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t \quad (2)$$

onde $\beta = (\beta_1, \ldots, \beta_k)^T$, k < n é um vetor de parâmetros de regressão desconhecidos $(\beta \in R^k)$, $x_{t1}, \ldots, x_{tk}$ são observações das *k* covariáveis conhecidas e fixadas e $g(.)$ é uma função monótona e duas vezes diferenciável, restrita ao intervalo (0,1), denominada função de

ligação. Nota-se que a variância de y_t é uma função de μ_t e, por conseguinte, dos valores das covariáveis. Logo, este modelo admite variável resposta com variância não constante.

Dentre as várias funções de ligação existentes a que utilizaremos neste trabalho pela sua difusão é a ligação *logit* descrita da seguinte forma

$$\mu_t = \frac{e^{x_t^T \beta}}{1 + e^{x_t^T \beta}},$$

onde $x_t^T = (x_{t1}, \ldots, x_{tk}),\ t = 1, \ldots, n$. Neste caso, o vetor de parâmetros de regressão tem uma fácil interpretação em termos das razões de chance. Para a estimação dos parâmetros do modelo Beta, consideramos, tal como no artigo de Ferrari e Cribari-Neto (2004), o método da máxima verossimilhança.

3. Critérios de Seleção de Modelos

Como já foi mencionado, os critérios de seleção de modelos são um guia para a escolha do melhor modelo.

Muitos pesquisadores assumem que o modelo verdadeiro existe, tem dimensão finita e que o mesmo pertence ao conjunto de modelos candidatos. Sob esta suposição, o objetivo da seleção do modelo é escolher o modelo verdadeiro a partir deste conjunto. Outros pesquisadores não aceitam a suposição supracitada, e assumem que ou o modelo verdadeiro tem dimensão infinita ou não está incluído no conjunto de modelos candidatos. O objetivo dos critérios de seleção do modelo é escolher um modelo que mais se aproxime do modelo verdadeiro a partir de um conjunto de modelos candidatos com dimensão finita. O modelo candidato que mais se aproxima do modelo verdadeiro é definido como o modelo adequado (SHIBATA, 1980).

Neste trabalho, consideramos a suposição da existência de um modelo verdadeiro de dimensão finita.

Apresentamos, a seguir, a definição de cada um dos critérios para o modelo Beta, destacando que a obtenção de cada um deles requer cálculos numéricos, que foram implementados computacionalmente para a realização deste trabalho.

Para escolher o melhor modelo, foram criados e apresentados na literatura vários critérios ao longo do tempo. O primeiro foi o coeficiente de determinação (R^2); contudo, sua escolha não é uma boa estratégia, pois o mesmo sempre aumenta com a inclusão de novas

covariáveis. Para contornar este problema, foi criado um coeficiente de determinação ajustado, denominado pseudo R^2 (R_p^2), que é definido como o quadrado do coeficiente de correlação amostral entre $\hat{\eta}$ e g(y). Note que $0 \leq R_p^2 \leq 1$ e quando $R_p^2 = 1$ existe uma concordância perfeita entre $\hat{\eta}$ e g(y) e, por consequencia, entre $\hat{\mu}$ e y.

Embora seja conhecido que medidas de tipo R_p^2 são viesadas e seu uso requer cautela (ver, por exemplo, WILLET e SINGER, 1988, RICCI e MARTINEZ, 2008), ele é uma ferramenta suporte em muitos estudos de modelagem com regressão. Dentre os possíveis modelos propostos, o melhor modelo é aquele que maximiza o pseudo R^2. Para o cálculo do mesmo utilizaremos a seguinte expressão:

$$R_p^2 = 1 - \frac{SSE/(n-k)}{SST/(n-1)}, \quad (3)$$

onde $SSE = \sum_{t=1}^{n} (g(y_t) - \hat{\eta}_t)^2$, $SST = \sum_{t=1}^{n} (g(y_t) - \overline{g(y_t)})^2$, n é o número de observações, *k* é o número de parâmetros do modelo proposto e $\sum_{t=1}^{n} (g(y_t)/n = \overline{g(y_t)}$.

Akaike (1974) propõe o critério AIC (*Akaike Information Criterion*), que foi desenvolvido a partir dos Estimadores de Máxima Verossimilhança para decidir qual é o modelo mais adequado, quando se utilizam muitos modelos com quantidades diferentes de parâmetros. Isto é, seleciona um modelo que esteja bem ajustado com um número reduzido de parâmetros. É provado que este critério é assintoticamente eficiente, no entanto, não é assintoticamente consistente. O AIC foi o primeiro critério baseado na informação de Kullback-Leibler (K-L), ele é assintoticamente não viesado para K-L. O critério AIC supõe que o modelo verdadeiro pertence ao conjunto de modelos ajustados. Esta suposição pode ser irrealista na prática, porém permite calcular valores esperados em distribuições centrais e considerar o conceito de sobreajustamento. Em geral,

$$AIC = -2l(\hat{\mu}, \hat{\phi}) + 2(k+1), \quad (4)$$

onde $l(\hat{\mu}, \hat{\phi})$ é o logaritmo da função de verossimilhança do modelo sob investigação avalida nas respectivas estimativas de máxima verossimilhança dos parâmetros μ e ϕ. Visto não ser o AIC adequado para pequenas amostras, Sugiura (1978) e Hurvich e Tsai (1989) derivaram o AICc estimando a discrepância esperada de Kullback-Leibler diretamente dos modelos de regressão. Hurvich e Tsai também adotaram a suposição de que o modelo verdadeiro pertence ao conjunto de modelos candidatos, e mostraram que o AICc de fato tem melhor

desempenho que o AIC em pequenas amostras, pois corrige o sobreajustamento (HURVICH e TSAI, 1989); porém é assintoticamente equivalente ao AIC e, portanto, é assintoticamente eficiente. Como no caso do AIC, os parâmetros associados ao modelo candidato são estimados por Máxima Verossimilhança. Temos que

$$AICc = -2l(\hat{\mu}, \hat{\phi}) + 2(k+1)(n/n-k-2). \quad (5)$$

Akaike (1978) e Schwarz (1978) introduziram critérios de seleção de modelos concebidos através de uma perspectiva Bayesiana. Schwarz desenvolveu o SIC (*Schwarz Information Criterion*) para seleção de modelos da família Koopman-Darmois, ao passo que Akaike desenvolveu o critério de seleção de modelos BIC (*Bayesian Information Criterion*) para seleção de modelos em regressão linear. Neste trabalho, utilizamos o critério BIC por se tratar de modelos de regressão. Ao contrário do AIC, o critério BIC assume que o modelo verdadeiro tem dimensão infinita e, portanto, não pertence ao conjunto de modelos candidatos. O BIC é definido por

$$BIC = -2l(\hat{\mu}, \hat{\phi}) + (k+1)\log(n). \quad (6)$$

Como é sinalizado em McQuarrie e Tsai (1998), o termo $2(k+1)\left(\frac{n}{n-k-2}\right)$ substituído por $(k+1)\log(n)$ resulta numa penalidade maior para o sobreajustamento. Hannan e Quinn (1979) propuseram o critério HQ para modelos de séries temporais autorregressivos, porém o mesmo pode ser estendido para outros modelos. Ele é obtido através de

$$BIC = -2l(\hat{\mu}, \hat{\phi}) + 2(k+1)\log(\log(n)). \quad (7)$$

Tanto o BIC quanto o HQ são assintoticamente consistentes, no entanto, muitos autores apontam que o HQ comporta-se como o critério eficiente AIC. Outro critério de seleção de modelos utilizado foi o HQc que se originou na modificação da função de penalidade do HQ, usando análise semelhante a que foi realizada entre o AIC e AICc, sendo definido por

$$HQc = -2l(\hat{\mu}, \hat{\phi}) + \frac{2(k+1)\log(\log(b))n}{n-k-3}. \quad (8)$$

Pode-se provar que o HQc não apenas corrige o desempenho do HQ em pequenas amostras em relação ao sobreajustamento, como também é um critério assintoticamente consistente.

Existem várias maneiras de comparar os diferentes critérios existentes na literatura. Uma maneira é analisar o número de vezes que cada critério identifica o verdadeiro modelo e uma outra maneira é a utilização de alguma medida de distância entre o modelo escolhido e o

verdadeiro modelo. Em ambos os casos, existe a necessidade de realizar simulações e também que o modelo verdadeiro pertença ao conjunto de modelos candidatos. Notando que em qualquer conjunto de modelos existirá algum modelo mais próximo ao verdadeiro modelo. A razão que utilizaremos que compara as distâncias entre os modelos escolhidos e o modelo mais próximo é chamada de eficiência observada, que será apresentada a seguir.

Lembrando que dado um vetor $x=(x_1,\ldots,x_n)$, o quadrado de sua norma euclideana é dado por $\|x\|^2=\sum_{i=1}^{n}x_i^2$ assim baseada nesta norma, a medida de distância utilizada neste trabalho L_2 é definida por

$$L_2=\|\mu(M_v)-\mu(M_c)\|^2, \quad (9)$$

onde $\mu(M_v)$ e $\mu(M_c)$, denotam o vetor de médias do modelo verdadeiro (M_v) e do modelo candidato (M_c). Uma vantagem de L_2 é que o mesmo só depende das médias das duas distribuições e não das densidades. Assim, a medida L_2 pode ser aplicada quando os erros não são normalmente distribuídos.

Shibata (1980) sugere o uso da distância esperada,

$$E(L_2)=E(\|\mu(M_v)-\mu(M_c)\|^2), \quad (10)$$

como medida de distância entre o modelo verdadeiro M_v e o modelo M_c. Usando esta medida, McQuarrie e Tsai (1998) assumem que existe entre os modelos candidatos um modelo que é o mais "aproximado" do modelo verdadeiro (M_a) em termos da esperança de L_2, isto é, que minimize (10). Isto é se

$$E(L_2(M_a))=E(L_2(M_a)=\|\mu(M_v)-\hat{\mu}(M_a)\|^2),$$

em que $\hat{\mu}$ é o vetor de valores preditos do modelo M_a e suponhamos que um critério de seleção de modelos escolha um determinado modelo M_k, tal que $E(L_2(M_k))=E(\|\mu(M_v)-\hat{\mu}(M_k)\|^2)$, temos que $E(L_2(M_k))\geq E(L_2(M_a))$.

Definimos a eficiência observada L_2

$$EOL_2=\frac{L_2(M_a)}{L_2(M_K)}, \quad (11)$$

Logo, o desempenho de um critério de seleção de modelos será melhor quanto maior seja sua eficiência observada.

4. Estudos de Simulação

Neste capítulo, analisamos dois particulares modelos Beta com o objetivo de verificar qual o melhor critério de seleção nestes casos, levando em consideração seis critérios de seleção de modelos (R2p , AIC, AICc, BIC, HQ, HQc) definidos nas equações (3) a (8). Destacando que dentre os critérios disponíveis na literatura houve uma preocupação de selecionarmos critérios assintoticamente eficientes e consistentes.

Avaliamos através de simulações de Monte Carlo o desempenho dos critérios para dois modelos diferentes (Modelo 1 e Modelo 2), levando em consideração diferentes tamanhos de amostras (n = 20; 40; 60; 200) e o número de réplicas R = 5000: Para a construção dos modelos, foram utilizadas sete covariáveis com distribuição exponencial de parâmetro 3 e incluímos o intercepto. Os valores de x_{tj} são gerados em cada réplica independentes e identicamente distribuídos segundo uma exponencial de parâmetro 3, sendo $t = 1,2,\ldots,n$ e $j = 1,2,\ldots,7$ e $x_{t0} = 1$. Obtendo em cada simulação um conjunto de 255 subconjuntos considerados como modelos candidatos potenciais, segundo a matriz de regressores tenha C colunas, com $C = 1,\ldots,8$. Não foram consideradas as interações. Por exemplo, um dos possíveis modelos de duas colunas é dado por $[x_{t0}, x_{t1}]$; num outro modelo também de duas colunas é $[x_{t0}, x_{t2}]$, totalizando assim 28 modelos possíveis com duas (ou 6) colunas. Da mesma forma, temos 56 modelos com três (ou 5) colunas; 70 modelos com quatro colunas; 8 modelos com uma (ou 7) coluna(s) e um único modelo com uma coluna, totalizando 255 modelos.

Para ambos, os modelos verdadeiros foram considerados o número total de parâmetros igual a 5, sendo o intercepto $\beta_0 = 1$.

Modelo 1:

O modelo verdadeiro será aquele em que

$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$.

Logo os valores de $y_t(y_1, \ldots, y_n)$ em cada réplica são gerados da distribuição $Beta(\mu_t, \phi^*)$, com ϕ^* = 120, onde

$$\mu_t = \frac{\exp(\beta_0 + \beta_1 x_{t1} + \ldots + \beta_4 x_{t4})}{1 + \exp(\beta_0 + \beta_1 x_{t1} + \ldots + \beta_4 x_{t4})}.$$

Posteriormente, calculamos os valores da distância L_2 por (9) as respectivas eficiências observadas L_2 por (11) de todos os modelos candidatos. Notemos que nesta situação o modelo verdadeiro é conhecido e, portanto, é possível calcular a distância L_2 e determinar assim dentre todos os modelos candidatos ajustados aquele que torna mínima esta distância, destacamos também que nem sempre esta distância seleciona o verdadeiro modelo. Em cada réplica, foi escolhido um modelo através dos critérios utilizados e a partir do modelo selecionado foi calculada a eficiência observada definida em (11). Posteriormente, calculamos a média, a mediana, o desvio padrão e o coeficiente de variação para as eficiências observadas resultantes de cada critério.

Modelo 2: é idêntico ao Modelo 1 com exceção dos valores dos parâmetros de β que são bem menores, definidos como

$\beta_1 = 1, \beta_2 = 1/2, \beta_3 = 1, \beta_4 = 1/4$.

A diminuição dos valores dos parâmetros de β, tal como referido por McQuarrie e Tsai (1998), torna o modelo “menos identificável" visto que as variáveis associadas a estes parâmetros têm uma menor associação com a variável resposta.

Contamos o número de vezes em que o critério selecionou o modelo verdadeiro, e o número de vezes em que os modelos de ordem C foram escolhidos, lembrando que C é o número de colunas da matriz de regressores. Posteriormente, calculamos a média, a mediana, o desvio padrão e o coeficiente de variação das eficiências observadas L_2. O desempenho dos critérios de seleção de modelos foi baseado na média da eficiência observada L_2, onde a maior eficiência observada denota melhor desempenho. Classificamos como tendo escore 1 (o melhor) o critério com a maior eficiência observada, ao passo que o critério com a menor eficiência observada foi classificado de escore 6 (o pior).

Os resultados das simulações para os dois modelos estão apresentados nas Tabelas A.1 a A.8 inseridas no Apêndice A.

Para o Modelo 1, a partir da Tabela A.1 com n = 20, considerando o número de vezes em que o critério escolheu o verdadeiro modelo M_v, podemos notar que o AICc dentre todos os critérios foi o que mais selecionou o modelo verdadeiro, obtendo uma frequência (2691) seguido pelo critério HQc (2628). Observamos, ainda, que critérios AIC e HQ tendem a sobreajustar, já que selecionam modelos de dimensão maior um número considerável de vezes. Este problema é corrigido, em parte, pelos critérios AICc e HQc. Podemos notar que as dispersões relativas da eficiência observada em todos os critérios estão muito próximas, porém os coeficientes de variação são consideravelmente altos, superiores a 42%. Podemos observar que o critério com melhor desempenho foi o AICc com eficiência observada mediana de 0,9842, seguido do BIC (0,7734) e do HQc (0,6993). O critério que obteve menor desempenho foi o R2p com eficiência observada média de 0,6548.

Na Tabela A.2 com n = 40, também observamos que o critério com melhor desempenho foi o HQc com eficiência observada média de 0,8930, seguido do BIC com 0,8771, ambos com uma eficiência mediana igual a 1. O menor desempenho foi do R2p com eficiência observada média de 0,7029.

Verificamos nas Tabelas A.3 e A.4 com n = 60 e n = 200, respectivamente, que nenhum dos critérios apresentou problemas de baixo-ajustamento e que o BIC foi o critério com melhor desempenho. Da comparação entre as Tabelas A.1 a A.4, observamos que, em geral, como era esperado, o número de vezes que o modelo escolhido é o verdadeiro, aumentou em todos os casos com o aumento de tamanho da amostra.

Também observamos que os critérios AIC e AICc têm o mesmo problema de sobreajustamento. Porém, o critério HQc parece corrigir este problema do HQ. Notamos, ainda, que em todos os tamanhos de amostra os piores critérios foram o R2p e o AIC. Em geral, para este modelo podemos concluir que o critério BIC obteve um bom desempenho, já que seu escore foi 1 ou 2 em todos os casos.

Para o Modelo 2, com os resultados apresentados nas Tabelas A.5 (n = 20) e A.6 (n = 40), destacamos que para este modelo as correções não funcionaram bem para os critérios AIC e HQ. Nestes casos, pode-se considerar que os critérios AIC e HQ tiveram um desempenho levemente superior. Para n = 60, Tabela A.7, percebemos que o AICc realiza correções no AIC, tendo o melhor desempenho, o mesmo não acontece com o HQc em relação ao HQ. Com as simulações realizadas para o Modelo 2, concluímos que não podemos eleger um critério como melhor, devido os mesmos assumirem escores diferentes

dependendo do tamanho da amostra. Na Tabela A.8 (n = 200), todos os critérios têm um desempenho similar, à exceção do BIC, que tem um desempenho levemente superior.

Comparando os resultados obtidos nas simulações para os dois tipos de modelos, confirmamos que o Modelo 1 representa o caso onde é mais fácil identificar como modelo candidato o modelo verdadeiro. Assim, da comparação das Tabelas do Modelo 1 (A.1 a A.3) com as do Modelo 2 (A.5 a A.7), respectivamente, podemos concluir que o número de vezes que o modelo verdadeiro é escolhido é bem menor para o Modelo 2 que para o Modelo 1. Também as eficiências médias e medianas do segundo modelo são menores, isto se deve, conforme já referido, ao fato de que o Modelo 2 é “fracamente identificável".

Vale destacar que, para n = 200, Tabelas A.4 e A.8, os dois Modelos apresentaram resultados bem próximos em todos os critérios aqui analisados e todos os critérios obedeceram à mesma ordem de classificação.

5. Aplicação

A aplicação foi feita com objetivo de escolher o melhor modelo para explicar o Índice de Desenvolvimento Humano Municipal - IDHM dos municípios nordestinos no ano de 2000, segundo IDHM do ano de 1991 e de outras características socioeconômicas, em cada um dos estados da Região Nordeste. O IDHM foi criado a partir do Programa das Nações Unidas para o Desenvolvimento - PNUD, que tem como objetivo central o combate à pobreza. O Índice de Desenvolvimento Humano Municipal - IDHM, para o Brasil, pode ser consultado no Atlas de Desenvolvimento Humano, um banco eletrônico com informações sócio-econômicas sobre os 5 507 municípios existentes no País no ano de 1991, os 26 Estados e o Distrito Federal e está baseado nos microdados dos Censos 1991 e 2000 do Instituto Brasileiro de Geografia e Estatística – IBGE.[5]

O IDH foi criado a partir dos seguintes indicadores: o de educação (alfabetização e taxa de matrículas), o de longevidade (esperança de vida ao nascer) e o de renda (Produto Interno Bruto *per capita*), com o intuito de medir o nível de desenvolvimento dos países. Em função da forma como é definido, o índice varia de 0 (nenhum desenvolvimento humano) a 1 (desenvolvimento humano total). Vale destacar que conceitualmente se considera de baixo desenvolvimento humano os países com IDH igual ou inferior a 0,499, de médio desenvolvimento, os que têm IDH entre 0,500 e 0,799 e, finalmente, de alto desenvolvimento, aqueles que detêm IDH maior ou igual a 0,800.

[5] **P**ara maiores informações ver http//www.undp.org.br. Segundo o IBGE, existem 5 564 municípios brasileiros instalados no Território Nacional até 31 de dezembro de 2007.

Após calcular os sub-índices: IDHM-E, para educação; IDHM-L, para saúde (ou longevidade); IDHM-R, para renda; obtém-se o IDHM, que é a média aritmética simples desses três sub-índices, ou seja;

$$\text{IDHM} = \frac{\text{IDHM - E} + \text{IDHM - L} + \text{IDHM - R}}{3}.$$

Como possíveis variáveis explicativas do IDHM-2000 foram consideradas as seguintes variáveis, cujos valores observados também foram obtidos do referido Atlas:

x_{t1}: IDHM de 1991;

x_{t2} : índice de Gini (GINI);

x_{t3}: proporção de indigentes (INDIG);

x_{t4} : proporção de pessoas que vivem em domicílios com água encanada (ÁGUA); e

x_{t5}: proporção de pessoas que vivem em domicílios urbanos com serviço de coleta de lixo (LIXO).

x_{t6} : proporção de pessoas que vivem em domicílios com energia elétrica (ENERGIA).

Ao se considerar as seis regressoras mais o intercepto, consideramos 127 modelos candidatos potenciais segundo a matriz de regressores que tenha C colunas, com $C = 1,\ldots,7$. Também nesta aplicação não forão consideradas as interações. Por exemplo, um dos possíveis modelos de duas colunas é dado por $[x_{t0}, x_{t1}]$ num outro modelo também de duas colunas é $[x_{t0}, x_{t2}]$, totalizando assim 21 modelos possíveis com duas (ou 5) colunas. Da mesma forma, temos 35 modelos com três (ou 4) colunas; 7 modelos com uma (ou 6) coluna(s) e um único modelo com uma coluna, resultando 127 modelos. Rotulamos todos os modelos candidatos de 1 a 127 com a finalidade de apresentar os resultados obtidos para cada estado da Região Nordeste. A diferença dos exemplos das simulações, o "verdadeiro modelo" é desconhecido.

A Tabela 1 mostra, para cada estado e para cada critério de seleção de modelo, os rótulos dos modelos selecionados, onde podemos observar que nos Estados de Pernambuco, Rio Grande do Norte e Alagoas todos os critérios selecionaram um único modelo para cada caso.

Os modelos selecionados, pelo menos por algum critério para algum estado, contêm o intercepto e a covariável IDHM de 1991. Sem dúvida, o fato da variável IDHM-1991 ter sido

selecionada em todos os modelos revela que a situação anterior (9 anos antes) do nível do IDHM é explicativa do nível do IDHM-2000, pois existe naturalmnte uma correlação entre o IDHM-1991 e o IDHM-2000. A especificação das demais covariáveis que compõe cada modelo rotulado é a seguinte:

Modelo 15: INDIG;
Modelo 16: INDIG. e ÁGUA;
Modelo 31: GINI e ENERGIA;
Modelo 33: GINI, INDIG, LIXO e ENERGIA;
Modelo 87: INDIG e ENERGIA;
Modelo 100: Não contém outras covariáveis, além do IDHM de 1991;
Modelo 102: GINI e INDIG;
Modelo 103: GINI, INDIG e ÁGUA; e
Modelo 104: GINI, INDIG, ÁGUA e LIXO.

Tabela 1: Rótulos dos modelos selecionados ordenados segundo o número de municípios (n) e o número de modelos diferentes escolhidos (NM)

Estados	Critérios						
	AIC	AICc	BIC	HQ	HQc	n	NM
Pernambuco	87	87	87	87	87	185	1
Rio Grande do Norte	103	103	103	103	103	166	1
Alagoas	100	100	100	100	100	101	1
Paraíba	16	16	15	15	15	223	2
Piauí	103	103	102	103	103	221	2
Maranhão	33	102	102	102	102	217	2
Ceará	31	31	102	102	102	184	2
Bahia	104	104	102	31	31	415	3
Sergipe	103	103	15	102	15	75	3

Os resultados encontrados em cada estado da Região Nordeste, apresentados na Tabela 1, revelam que os diversos critérios não selecionaram, em geral, o mesmo modelo. Constatamos que os critérios AIC e AICc escolheram os modelos diferentes unicamente no Estado do Maranhão, o AIC escolheu um modelo com duas variáveis a mais, mostrando talvez um problema de sobreajustamento por parte do AIC; já que o mesmo modelo foi escolhido pelos outros critérios também. No Estado da Paraíba, os critérios AIC e AICc selecionaram o modelo 16, que contém uma variável a mais que o modelo 15 selecionado pelos demais critérios, que tem uma penalidade maior ao sobreajustamento.

No Estado de Ceará, critérios AIC e AICc escolheram o modelo 31, enquanto os demais critérios selecionaram o modelo 102. Estes modelos se diferenciam nas variáveis ENERGIA e

INDIG, contidas num e não no outro modelo. Para o Estado do Piauí, o critério BIC escolheu um modelo de dimensão menor, isto é, com uma variável a menos, que os outros critérios.

Também poderia se suspeitar de problemas de sobreajustamento por parte de AIC, AICc e HQ nos modelos escolhidos para o Estado de Sergipe, dado que HQc escolheu o modelo 15 contido nos modelos 103 e 102 escolhidos por estes critérios, respectivamente. No Estado da Bahia, os critérios AIC, AICc selecionaram o modelo 104 contido no modelo 102 selecionado pelo BIC, enquanto os demais critérios escolheram o modelo 31.

Destacamos que dado que modelos com interações não foram considerados, não é possível tirar conclusões definitivas desta aplicação que teve por objetivo simplesmente analisar possíveis situações de sobreajustamento como foram observadas nas simulações. Os critérios apresentados indicam o caminho inicial da procura do modelo mais adequado, sendo necessária uma análise detalhada da significância dos parâmetros associados a cada modelo e das análises de resíduos e de diagnóstico correspondentes.

6. Conclusões

A partir dos resultados das simulações, constatamos que para o Modelo 1, “mais identificável", o critério BIC sempre obteve escores 1 e 2, logo poderia ser considerado como o melhor.

Constatamos que num modelo “menos identificável" (Modelo 2) não foi possível eleger o melhor critério, confirmando os achados de McQuarrie e Tsai (1998) de que um critério nem sempre é melhor que outro, pois depende do modelo verdadeiro.

Confirmamos, ainda, que o número de vezes que o modelo verdadeiro é selecionado aumenta quando a amostra aumenta, independente do modelo verdadeiro.

Vale destacar que, para n = 200 Tabelas A.4 e A.8, os dois Modelos apresentam resultados bem próximos em todos os critérios aqui analisados e todos os critérios obedeceram à mesma ordem de classificação, o que nos leva a conclusão que quando o tamanho da amostra é suficientemente grande a ordem de classificação dos critérios independe do modelo verdadeiro.

A aplicação teve um caráter ilustrativo em termos de desempenho dos critérios em relação a sobreajustamento. Porém, destacamos a importância das variáveis: IDHM de 1991, percentual de indigentes e Índice de Gini que surgem da comparação dos modelos escolhidos pelos diferentes critérios para cada um dos estados.

Apesar de não ser simples a tarefa de determinar qual é o critério de melhor desempenho, dado que os critérios de seleção de modelos são muito importantes porque indicam o caminho inicial para escolha do modelo mais apropriado a um conjunto de dados vale a pena continuar a pesquisa neste assunto analisando que tipo de relações entre as variáveis de um modelo levam a uma falta de uniformidade entre os critérios.

Referências bibliográficas

[1] Akaike, H. (1974). A new look at statistical model identification. IEEE. Transactions on Automatic Control AU, 19, 716-722.

[2] Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics A, 30, 9-14.

[3] Ferrari, S. L. P. e Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. Journal of Applied Statistics, 31, 7, 799-816.

[4] Hannan, E. J. e Quinn, B. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society B, 41, 190-191.

[5] Huvich, C. M. e Tsai, C-L. (1989). Regression and time series model selection in small samples. Biometrics, 76, 297-307.

[6] Kuha, J. (2004). AIC and BIC - Comparisons of assumptions and performance. Sociological Methods & Research, 33, 2, 188-229.

[7] Mallows, C. L. (1973). Some comments on Cp. Technometrics, 37, 661-675.

[8] McQuarrie, A. D. R. e Tsai, C-L. (1998). Regression and Time Series Model Selection. Singapore: World Scientiç Publishing Co. Pte. Ltd.

[9] Nações Unidas (2003). PNUD - Programa das Nações Unidas para o Desenvolvimento: Atlas do Desenvolvimento Humano no Brasil. http//www.undp.org.br.

[10] Rao, C. R. e Wu, Y. (2005). Linear Model Selection by Cross-validation. Journal of Statistical Planning and Inference, 128, 1, 231-240.

[11] Ricci, L e Martínez, R. (2008). Adjusted R2 type measure for Tweedie models. Computational Statistics & Data Analysis, 52, 1650-1660.

[12] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

[13] Shibata, R. (1980). An optimal selection of regression variables. Biometrika, 68, 45-54.

[14] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. Communicationin Statistics – Theory and Methods, 7, 13-26.

[15] Willet, J.B e Singer, J.D. (1988). Another cautionary note about R2: Its use in weighted least-squares regression analysis. Journal of the American Statistical Association, 42, 3, 236-238.

Abstract

The Beta regression model holds a great practical applicability, in particular, for modelling rates and proportions and such as the others regression models, it also requires methods to determinate which is the best model. The main objetive of this work is to implement and evaluate the performance of diérents model selection criteria in the Beta regression model. For such, by using diérents studies of Monte Carlo simulations, we have analised some criteria selected by taking into consideration its asymptotic properties, which were obtained by maximum likelihood function. The simulations results show that the performances of those criteria depend on the model speciçation as well as on the sample size. We have also presented an application related to the Human Development Index from United Nations Development Programme (UNDP), which is a right variable for the modelling in study, since its values vary in the interval (0,1).

Agradecimentos

Os autores agradecem o apoio financeiro parcial da Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP. Os autores também agradecem os comentários e sugestões dos pareceristas e editores.

Apêndice A

Tabela A.1: Frequência do modelo selecionado e as eficiências observadas para o Modelo 1 com *n* = 20; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	2	2	0	5	0	0
2	1	21	9	1	44	0	0
3	29	205	100	42	329	23	0
4	256	977	532	302	1249	260	28
5	2063	3051	2642	2227	2949	1781	3681
6	1751	677	1317	1660	397	1916	1136
7	749	64	343	649	26	860	147
8	151	3	55	119	1	160	8
verdadeiro	1793	2691	2318	1939	2628	1368	3633

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,6816	0,7317	0,7135	0,6911	0,7134	0,6413	1,0000
Mediana	0,6871	0,9842	0,7734	0,6993	0,9796	0,6548	1,0000
Desvio-padrão	0,2881	0,3152	0,3009	0,2911	0,3270	0,2992	0,0000
C.V.(%)	42,27	43,08	42,17	42,12	45,84	46,66	0,00
Escore	5	1	2	4	3	6	

Tabela A.2: Frequência do modelo selecionado e as eficiências observadas para o Modelo 1 com *n* = 40; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	5	11	26	7	26	42	0
5	2599	3522	3552	3237	4102	1818	4066
6	1820	1277	905	1455	805	2132	846
7	520	180	110	276	64	882	84
8	56	10	7	25	3	126	4
verdadeiro	2593	3514	3942	3231	4092	1743	4066

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,7578	0,8375	0,8771	0,8125	0,8930	0,7029	1,0000
Mediana	0,9889	1,0000	1,0000	1,0000	1,0000	0,7526	1,0000
Desvio-padrão	0,2850	0,2687	0,2495	0,2769	0,2373	0,2904	0,0000
C.V.(%)	37,61	32,09	28,45	34,08	26,57	41,31	0,00
Escore	5	3	2	4	1	6	

Tabela A.3: Frequência do modelo selecionado e as eficiências observadas para o Modelo 1 com *n* = 60; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	21	0
5	2662	3227	4206	3446	4024	1893	4219
6	1844	1524	731	1332	892	2133	725
7	455	241	61	215	81	839	55
8	39	8	2	7	3	114	1
verdadeiro	2662	3227	4205	3446	4023	1850	4219

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,7526	0,8058	0,9034	0,8259	0,8838	0,7035	1,0000
Mediana	0,9955	1,0000	1,0000	1,0000	1,0000	0,7521	1,0000
Desvio-padrão	0,2934	0,2835	0,2307	0,2774	0,2464	0,2932	0,0000
C.V.(%)	38,98	35,18	25,54	33,59	27,88	41,68	0,00
Escore	5	4	1	3	2	6	

Tabela A.4: Frequência do modelo selecionado e as eficiências observadas para o Modelo 1 com *n* = 200; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	8	0
5	2953	3124	4633	3966	4116	2064	4519
6	1678	1574	357	964	840	2112	465
7	345	286	10	67	41	730	16
8	24	16	0	3	3	86	0
verdadeiro	2953	3124	4633	3966	4116	2052	4519

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,7778	0,7933	0,9501	0,8751	0,8917	0,7245	1,0000
Mediana	1,0000	1,0000	1,0000	1,0000	1,0000	0,7860	1,0000
Desvio-padrão	0,2933	0,2906	0,1817	0,2568	0,2442	0,2879	0,0000
C.V.(%)	37,71	36,33	19,12	29,35	27,39	39,74	0,00
Escore	5	4	1	3	2	6	

Tabela A.5: Frequência do modelo selecionado e as eficiências observadas para o Modelo 2 com *n* = 20; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	11	10	2	17	0	0
2	91	370	278	121	573	24	6
3	538	1508	1060	651	1831	258	120
4	1397	1980	1723	1482	1861	966	1011
5	1572	931	1264	1524	617	1720	3248
6	1003	181	507	882	90	1384	568
7	343	18	142	296	11	560	47
8	56	1	16	42	0	88	0
verdadeiro	552	394	492	552	283	547	2628

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,5682	0,5522	0,5595	0,5692	0,5355	0,5698	1,0000
Mediana	0,5436	0,5278	0,5353	0,5458	0,5076	0,5475	1,0000
Desvio-padrão	0,2529	0,2568	0,2571	0,2549	0,2536	0,2495	0,0000
C.V.(%)	45,51	46,50	45,95	44,78	47,36	43,79	0,00
Escore	3	6	4	2	5	1	

Tabela A.6: Frequência do modelo selecionado e as eficiências observadas para o Modelo 2 com *n* = 40; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	0	1	0	0	0	0
2	1	5	45	7	33	1	0
3	150	301	660	292	613	48	5
4	992	1569	1904	1417	2006	480	248
5	2239	2310	1929	2203	1996	1772	4173
6	1243	703	399	889	311	1824	539
7	337	107	61	176	41	750	33
8	38	5	1	16	0	125	2
verdadeiro	1538	1628	1439	1564	1481	1109	3994

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,6600	0,6606	0,6280	0,6550	0,6340	0,6393	1,0000
Mediana	0,6437	0,6415	0,5983	0,6319	0,6074	0,6343	1,0000
Desvio-padrão	0,2851	0,2929	0,2979	0,2906	0,2986	0,2698	0,0000
C.V.(%)	43,20	44,34	47,44	44,37	47,10	42,20	0,00
Escore	2	1	6	3	5	4	

Tabela A.7: Frequência do modelo selecionado e as eficiências observadas para o Modelo 2 com *n* = 60; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	0	0	0	0	0	0
2	0	1	10	1	2	0	0
3	34	58	277	87	165	7	1
4	568	831	1621	1023	1410	237	69
5	2561	2840	2631	2817	2825	1757	4410
6	1442	1090	436	925	560	1994	496
7	372	175	25	142	38	874	24
8	23	5	0	5	0	131	0
verdadeiro	2174	2438	2334	2421	2460	1421	4353

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,7188	0,7371	0,7073	0,7313	0,7269	0,6754	1,0000
Mediana	0,7384	0,8841	0,7707	0,8675	0,9372	0,6689	1,0000
Desvio-padrão	0,2855	0,2904	0,3083	0,2941	0,3022	0,2683	0,0000
C.V.(%)	39,72	39,40	43,59	40,22	41,57	39,72	0,00
Escore	4	1	5	2	3	6	

Tabela A.8: Frequência do modelo selecionado e as eficiências observadas para o Modelo 2 com *n* = 200; para *R* = 5000:

C	Frequência						
	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	5	6	87	21	25	2	0
5	2953	3099	4574	3999	4110	1707	4703
6	1676	1598	326	891	795	2237	288
7	340	278	13	84	70	907	9
8	26	19	0	5	0	147	0
verdadeiro	2951	3096	4567	3995	4103	1704	4703

Eficiência observada L^2

Medidas	AIC	AICc	BIC	HQ	HQc	R^2_p	L^2
Média	0,7912	0,8038	0,9452	0,8852	0,8966	0,7117	1,0000
Mediana	1,0000	1,0000	1,0000	1,0000	1,0000	0,7342	1,0000
Desvio-padrão	0,2772	0,2747	0,1831	0,2411	0,2324	0,2680	0,0000
C.V.(%)	35,04	34,18	19,37	27,24	25,92	37,66	0,00
Escore	5	4	1	3	2	6	

Aplicação do modelo de Cox para identificar fatores de risco em pacientes com câncer de mama

Cláudia Patrícia Costa de Macedo[6]
Dione Maria Valença[7]

Resumo

As técnicas estatísticas de análise de sobrevivência tratam os dados relativos ao tempo até a ocorrência de um fenômeno. Este artigo utiliza o modelo de análise de sobrevivência, o modelo de regressão de Cox (Cox,1972). A suposição de riscos proporcionais foi considerada no conjunto dos dados analisados. O câncer de mama é um grave problema de saúde pública no Brasil. E causa uma alta mortalidade entre as mulheres. Nosso objetivo principal foi verificar dentre as informações registradas das pacientes em estudo, ou seja, dentre algumas variáveis explanatórias, quais poderiam influenciar de forma significativa o tempo em que a paciente, com câncer de mama, permanecia livre do retorno (recidiva) da doença.

Palavras-chave: Análise de sobrevivência, modelo de regressão de Cox, câncer de mama.

[6] Departamento de Estatística, Centro de Ciências Exatas e da Terra, Universidade Federal do Rio Grande do Norte
[7] idem.

1. Introdução

Os estudos de sobrevivência são caracterizados pelos tempos até a ocorrência de um evento, isto é, o tempo transcorrido entre um evento inicial, que marca a entrada do indivíduo no estudo, até um evento final que modifica este estado inicial. Esse tempo é denominado tempo de falha ou tempo de vida.

Em dados de sobrevivência provenientes de estudos da área médica, frequentemente, o evento inicial pode corresponder ao diagnóstico de uma doença ou à remissão desta (momento em que o paciente fica livre da doença), e o evento final pode ser a morte do paciente ou a recidiva da doença (recorrência da doença). Nestes casos, o tempo de falha pode ser, por exemplo, o tempo decorrido do diagnóstico da doença até a morte do indivíduo ou mesmo o tempo da remissão até a recidiva da doença.

Em estudos, como esses, onde há seguimento no tempo, pode acontecer de alguns indivíduos não serem acompanhados até a ocorrência da falha, ou seja, o tempo de observação é parcial. Este tipo de perda na observação é denominado censura (BUSTAMANTE-TEIXEIRA *et al*., 2001). Na presença de censuras, faz-se necessário o uso de métodos de análise de sobrevivência que são capazes de incorporar a informação contida nos dados censurados.

Na maioria dos estudos médicos, são obtidas informações complementares de cada paciente. Por exemplo, em oncologia, podemos observar o tempo de recidiva do câncer de mama para as pacientes submetidas a um determinado tipo de tratamento cirúrgico para remoção completa do tumor. Contabilizamos, portanto, quanto tempo cada paciente ficou livre da doença. A distribuição deste tempo de recidiva da doença pode depender da idade da paciente no momento do diagnóstico da enfermidade, do tipo de tratamento cirúrgico ao qual a paciente foi submetida, do tamanho do tumor, entre outras características que irão dividir as pacientes em grupos distintos. Assim, para cada indivíduo da amostra estará associado um vetor contendo essas informações auxiliares. Segundo Maller e Zhou (1996) essas informações são chamadas de variáveis explanatórias ou variáveis regressoras ou covariáveis. Utilizando um modelo apropriado que incorpore essas informações na análise, podemos explorar como a ocorrência de um evento em um grupo de pacientes depende de uma ou mais covariáveis, cujos valores foram registrados para cada paciente no momento da sua entrada no estudo. O modelo de riscos proporcionais, proposto por Cox (1972), mais conhecido como modelo de regressão de Cox, é extensivamente utilizado em pesquisas medica e biológicas e representa um modelo que incorpora covariáveis para examinar o relacionamento destas com o risco de falha.

Em virtude da importância do câncer de mama, como um problema de saúde pública no Brasil (ALBERG, VISVANATHAN, e HELZLSOUER, 1998), esse trabalho se propõe a utilizar o modelo de regressão de Cox para identificar fatores capazes de influenciar o tempo em que as pacientes pernamenem livres do retorno (recidiva) da doença, após terem sido submetidas ao tratamento cirúrgico de retirada total ou parcial da mama.

Este trabalho encontra-se estruturado conforme segue: na seção 2, é discutido o modelo de regressão de Cox, incluindo o ajuste do modelo, comparação entre modelos, a estatística $\hat{L}$ ou -2log$\hat{L}$, e as estratégias para seleção do modelo. Na seção 3, é apresentada a aplicação do modelo de regressão de Cox na análise de dados de pacientes com câncer de mama. Na seção 4, são expostos os resultados da análise dos dados; e na última seção, são descritas as conclusões.

2. Modelo de Regressão de Cox

Considere uma amostra aleatória de *n* indivíduos e sejam $\mathbf{x}_i^T = (x_{i1},x_{i2},....,x_{ik})$ e h_i, respectivamente, o vetor (transposto) de covariáveis (ou variáveis explanatórias) e a função de risco associada ao indivíduo *i*. Assim, nossos dados consistem de *n* observações na forma $(t_i, \delta_i, \mathbf{x}_i)$, sendo t_i, δ_i e $\mathbf{x}_i$, respectivamente, o tempo de falha, o indicador de censura e o vetor de covariáveis. Seja $\beta = (\beta_1,\beta_2,....,\beta_k)$ um vetor de parâmetros desconhecidos e definido $\mathbf{x}_i^T$ como sendo o componente linear. A forma geral do modelo de regressão de Cox é dada por:

$$h_i(t) = \Psi(\mathbf{x}_i^T \beta)\, h_0(t) = \Psi(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})\, h_0(t),$$

sendo $\Psi(\mathbf{x}_i^T \beta)$ um componente paramétrico, não negativo, que envolve as covariáveis, mas não depende do tempo, conhecido como risco relativo ou razão de risco; $h0(t)$ um componente não paramétrico desconhecido. É uma função não negativa do tempo, mas não envolve as covariáveis. Usualmente denominada função de risco base ou função de risco padrão, pois $hi(t) = h0(t)$ quando o vetor $xi = 0$.

Em geral, é conveniente assumir para $\Psi(\mathbf{x}_i^T \beta)$ a seguinte forma

$$\Psi(\mathbf{x}_i^T \beta) = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik.}) \qquad (1)$$

O modelo de riscos proporcionais de Cox, para o i-ésimo indivíduo, pode ser escrito na forma:

$$h_i(t) = \exp\left\{\mathbf{x}_i^T \beta\right\} h_0(t), \quad i = 1, 2, ..., n. \qquad (2)$$

O modelo (2) é conhecido como Modelo de Riscos Proporcionais de Cox, pois considera a suposição de proporcionalidade entre as funções de risco. Para entender esta suposição, considere que a razão entre as funções de risco de falha de dois indivíduos *i* e *j* é dada por:

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp\left\{\mathbf{x}_i^T \beta\right\} h_0(t)}{\exp\left\{\mathbf{x}_j^T \beta\right\} h_0(t)} = \exp\left\{\mathbf{x}_i^T \beta - \mathbf{x}_j^T \beta\right\}. \qquad (3)$$

Assim, se no início do estudo, o risco de morte de um indivíduo é, por exemplo, duas vezes o risco de um outro indivíduo, esta razão de risco não depende do tempo, isto é, será a mesma durante todo o acompanhamento.

Com base em (3) o modelo dado em (2) pode ser expresso na seguinte forma:

$$log \left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta 1 xi1 + \beta 2 xi2 + ... + \beta k xik .$$

Podemos, então, dizer que o modelo de riscos proporcionais é um modelo linear para o logaritmo da razão de risco.

2.1. Ajuste do Modelo

O modelo de regressão de Cox é caracterizado pela inclusão de covariáveis que interferem na função de risco. Os coeficientes β's medem os efeitos das covariáveis sobre a função de risco (e consequentemente sobre a função de sobrevivência). O ajuste do modelo é determinado quando obtemos, a partir de observações amostrais, um vetor $\hat{\beta}$ de estimativas dos coeficientes β que são os parâmetros desconhecidos do modelo. Um método de estimação bastante conhecido é o método de máxima verossimilhança, no entanto a presença do componente não-paramétrico *h0*(t) no modelo torna este método inadequado. Cox (1972) no seu artigo original formalizado em um artigo subsequente (Cox, 1975), introduziu uma nova forma de função de verossimilhança denominada de verossimilhança parcial. O método proposto consiste em condicionar a verossimilhança para eliminar a função *h0*(t).

A função de verossimilhança parcial, $L(\beta)$, é dada pelo produto de todos os termos associados aos distintos tempos de falha e pode ser expressa na forma:

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(x_i^T \beta)}{\sum_{j \in R(t_{(i)})} \exp(x_j^T \beta)} = \prod_{i=1}^{n} \left[\frac{\exp(x_i^T \beta)}{\sum_{j \in R(t_{(i)})} \exp(x_j^T \beta)} \right]^{\delta_i}, \qquad (4)$$

sendo *xiT* o vetor de covariáveis observadas para o indivíduo que falhou no i-ésimo tempo de falha ordenado, e *δi* o indicador de censura. Do ponto de vista computacional, é mais conveniente maximizar o logaritmo da função de verossimilhança, que nesse caso é:

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i^T \beta - \log \sum_{j \in R(t_{(i)})} \exp(x_j^T \beta) \right].$$

Assim, a função de verossimilhança parcial apresenta duas características vantajosas: primeira, o componente não-paramétrico é eliminado e, segunda, a função não é afetada por tempos de vida censurados. Contudo, note que esta verossimilhança assume que os tempos são contínuos e, desta forma, não seria possível a ocorrência de tempos de sobrevivência empatados. Segundo Collet (1994), na literatura foram propostas modificações para a função de verossimilhança parcial de Cox no intuito de incorporar a presença de empates nos tempos de falha.

As estimativas de máxima verossimilhança para os parâmetros *β's* no modelo de riscos proporcionais, denotadas por, $\hat{\beta}$ podem ser obtidas através de métodos numéricos que buscam os valores que maximizam esta função. O erro padrão de cada uma das estimativas é denotado por e.p. ($\hat{\beta}$). Os valores das estimativas $\hat{\beta}$ são obtidos resolvendo o sistema de equações definido por *U(β) = 0*, onde *U(β),* chamado vetor escore, é o vetor das primeiras derivadas da função *ℓ(β) = log L(β).* Isto é:

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in R(t_{(i)})} x_j^T \exp(x_j^T \beta)}{\sum_{j \in R(t_{(i)})} \exp(x_j^T \beta)} \right].$$

As propriedades assintóticas dos estimadores de máxima verossimilhança são necessárias para construir intervalos de confiança e testar hipóteses sobre os coeficientes do modelo. Os autores Andersen (1982) e Gill e Schumacher (1987) mostraram que os estimadores obtidos pela maximização da verossimilhança parcial dada em (4) são consistentes e assintoticamente normais, sob certas condições de regularidade. Desta forma, podemos utilizar as conhecidas estatísticas de Wald e da razão de verossimilhança para fazer inferências no modelo de regressão de Cox.

2.2. Comparação entre Modelos

Um possível procedimento é o desenvolvimento de um modelo para verificar a dependência existente entre a função de risco e uma ou mais variáveis exploratórias. Na realização desse procedimento, são ajustados modelos de riscos proporcionais contendo diferentes termos no componente linear. Estes modelos são então comparados com o objetivo de determinar que termos são necessários na estrutura linear para uma razoável descrição dos dados.

Supondo que dois modelos são comparados para um grupo de dados em particular. O modelo 1 contém covariáveis *x1,x2,...,xp*, e função de risco da forma:

$$h_i(t) = \exp\left\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right\} h_0(t), \quad i = 1, 2, ..., n.$$

Um segundo modelo (modelo 2) contém *p+q* covariáveis *x1,x2,.....,xp,xp+1,xp+2,.....,xp+q*, ou seja, este modelo contém *q* variáveis explanatórias adicionais comparado ao modelo 1. Como o modelo 2 contém um número maior de covariáveis, ele deve se ajustar melhor aos dados observados. O problema estatístico será determinar se os *q* termos adicionais no modelo 2 melhoram significativamente o ajuste. Se não, o modelo 1, mais simples, será o mais adequado. Como já foi anteriormente comentado, quando existem várias variáveis explanatórias relevantes no estudo, o efeito de cada variável não pode ser estudado independentemente do efeito das outras. Por exemplo, no modelo 1 o efeito de qualquer uma das *p* covariáveis, na função de risco, depende das *p-1* variáveis anteriormente incluídas no modelo. Da mesma forma quando variáveis são adicionadas no modelo 1, o efeito dessas variáveis na função de risco é ajustado na presença das variáveis que já se encontravam no modelo.

2.3. A Estatística $\hat{L}$ ou *-2log* $\hat{L}$

É necessário utilizar uma estatística que venha mensurar a qualidade de ajuste de um modelo aos dados. Assim, a função de verossimilhança sumariza a informação que os dados contêm sobre os parâmetros desconhecidos num modelo. Uma estatística adequada é o valor da função de verossimilhança quando os parâmetros são substituídos por suas estimativas de máxima verossimilhança, denotada por $\hat{L}$. Isto é, a verossimilhança maximizada sob o modelo assumido. Quanto maior o valor da verossimilhança maximizada, melhor é a

concordância entre o modelo proposto e os dados observados. Como no nosso caso, $\hat{L}$ é de fato um produto de uma série de probabilidades condicionais, isto significa que esta estatística é menor que uma unidade, ou seja, 0< $\hat{L}$ <1. Por este motivo é mais conveniente utilizar *-2log* $\hat{L}$, que resultará num valor sempre positivo. O menor valor de *-2log* $\hat{L}$ significa um maior valor de $\hat{L}$ e, por conseguinte, o melhor modelo. Este valor só é útil em comparações entre modelos ajustados a um mesmo conjunto de dados. Esta comparação entre modelos é realizada pela diferença entre as estatísticas *-2log* $\hat{L}$ dos dois modelos. É usado o fato de que esta diferença possui assintoticamente uma distribuição qui-quadrado, sob a hipótese nula de que os coeficientes das variáveis adicionais são zero. O número de graus de liberdade é igual ao número de parâmetros que estão sendo acrescentados no modelo. Esta estatística será denominada de estatística da razão de verossimilhança.

Considere o modelo 1 com *p* variáveis e o modelo 2 com *p+q* variáveis. Denote por $\hat{L}_1$ e $\hat{L}_2$ os valores que maximizam a função para cada modelo, respectivamente. Os dois modelos podem ser comparados com base na diferença entre $-2\log\hat{L}_1$ e $-2\log\hat{L}_2$ (razão de verossimilhança). Esta diferença entre os valores $-2\log\hat{L}_1$ e $-2\log\hat{L}_2$ irá refletir o efeito combinado das *q* variáveis adicionais, ou seja, a alteração no valor de $-2\log\hat{L}$ devido ao acréscimo das *q* variáveis *xp+1,xp+2,.....,xp+q* ajustadas ao modelo com *x1,x2,...,xp*. A estatística da razão de verossimilhança ξ_{RV}, para testar a hipótese nula que os *q* parâmetros $\beta_{p+1},\beta_{p+2},....,\beta_{p+q}$ no modelo 2 são todos iguais a zero, pode ser expressa na forma:

$$\xi_{RV} = -2\log\hat{L}_1 - \left(-2\log\hat{L}_2\right) = -2\log\hat{L}_1 + 2\log\hat{L}_2 = -2\log\left\{\frac{\hat{L}_1}{\hat{L}_2}\right\}.$$

Pode-se mostrar que, sob *H0,* ξ_{RV} tem aproximadamente uma distribuição qui-quadrado com *q* graus de liberdade ($\xi_{RV} \sim \chi^2_q$).

2.4. Estratégias para seleção do modelo

O passo inicial na seleção do modelo é identificar quais as variáveis explanatórias com o poder potencial de explicar o risco de morte dos indivíduos de acordo com o objetivo principal do estudo. Essas variáveis serão incluídas no componente linear do modelo. Portanto, esse componente linear pode ser formado de fatores, variáveis e termos correspondentes a interações entre fatores, interações entre variáveis numéricas ou interações entre fatores e variáveis numéricas.

A estratégia de seleção depende dos propósitos do estudo. Uma etapa fundamental no processo de modelagem é avaliar o efeito de cada variável na função de risco ou função de sobrevivência. Usualmente os *softwares* estatísticos apresentam rotinas automáticas baseadas nos seguintes processos de seleção das variáveis explanatórias: seleção *forward*, eliminação *backward* e combinação das duas que é o procedimento *stepwise*. Cada procedimento de seleção apresentando suas peculiaridades (COLLETT, 1994).

Na prática, ao invés de utilizar esses procedimentos automáticos, alguns passos são recomendados como estratégia para seleção do modelo nas duas situações acima comentadas. Collet (1994) traz comentários com relação a cada tipo de procedimento automático e também enfoca os passos recomendados para uma estratégia não automática de seleção das variáveis explanatórias.

3. Aplicação do Modelo de Regressão de Cox

Nessa seção, apresentamos uma aplicação do modelo de regressão de Cox, através de um estudo retrospectivo de casos envolvendo pacientes com câncer de mama. Nosso intuito foi verificar, através do Modelo de Regressão de Cox, quais as variáveis que influenciam de forma significativa no tempo livre da doença em pacientes com câncer de mama. Para tanto, realizaremos etapas iniciais de seleção de variáveis e após esta fase apresentaremos os resultados com o modelo final proposto.

3.1. Descrição do Estudo

A coleta de dados foi realizada em prontuários médicos de 485 pacientes admitidos no hospital de referência estadual para neoplasias, Hospital Dr. Luiz Antônio. Esses pacientes tinham diagnóstico de câncer de mama comprovado através de exame anátomo-patológico, no período de 1991 a 1995. Após a etapa de verificação dos critérios de exclusão, permaneceram no estudo 355 mulheres com história de câncer de mama. Os principais motivos de exclusão foram: ausência de tratamento cirúrgico para retirada do tumor, diagnóstico inicial já com metástase a distância, data desconhecida do exame anátomo-patológico, data desconhecida do diagnóstico da recidiva, câncer de mama bilateral. Foram também eliminados os casos ocorridos em pacientes do sexo masculino.

O evento de interesse no nosso estudo foi o tempo decorrido entre a remissão e a recidiva do câncer de mama, ou seja, o tempo livre da doença. Assim, estes dados se caracterizam como dados de sobrevivência.

O mecanismo de censura presente nesse conjunto de dados foi a censura aleatória (perda de acompanhamento, pacientes que no final do estudo não apresentaram recidiva da doença, mudança de endereço, entre outros).

A suposição de riscos proporcionais foi considerada no conjunto dos dados analisados.

As variáveis explanatórias que entraram no estudo foram: idade (ID), tipo histológico do tumor (THIS), estadiamento do tumor (EST), proporção de linfonodos comprometidos (PLC), tipo de cirurgia a qual a paciente foi submetida para remoção do tumor (TC), tipo de tratamento não cirúrgico indicado (TNC) e tamanho do tumor em cm (TM, sendo considerada a dimensão de maior diâmetro). Na área médica, é muito comum utilizar a variável “número de linfonodos comprometidos”, propomos neste trabalho substituir essa variável pela “proporção de linfonodos comprometidos” que apresentou um melhor ajuste no modelo. Esta variável foi obtida através da razão entre número de linfonodos comprometidos pelo número de linfonodos ressecados. Apresentamos na Figura 1 a organização da pesquisa.

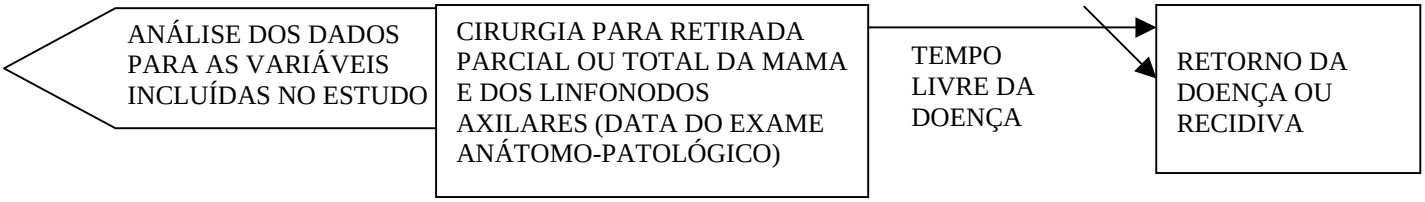


Figura 1. Esquema descrevendo organização do estudo

Podemos ver através da Figura 1 que inicialmente foram coletados os prontuários dos pacientes diagnosticados com câncer de mama no período de janeiro/1991 a dezembro/1995. Após essas etapas, foram analisados os dados para exclusão dos pacientes que não atendiam os critérios de inclusão do estudo. Através dos prontuários, observamos a data do exame anátomo-patológico para todas as pacientes incluídas no estudo. Esta data correspondia ao momento da cirurgia para retirada do tumor e dos linfonodos axilares. Através do exame anátomo-patológico o laboratório constatava a presença ou ausência de metástase local a nível de linfonodos. No caso da presença de metástase, era informado o número de linfonodos comprometidos dentre aqueles retirados. Considerando que após a cirurgia as pacientes podiam ser consideradas livres da doença, o tempo da data da cirurgia até a recidiva do câncer de mama refletia este tempo em que cada paciente permanecia sem a doença. As pacientes atendidas de janeiro de 1991 a dezembro de 1995 entraram no estudo e o tempo final do estudo foi fixado em dezembro de 2002 e a escala de tempo foi o tempo cronológico medido em meses.

A categorização das variáveis foi baseada na análise visual de gráficos de Kaplan-Meier e em trabalhos apresentados na área de câncer mama. As variáveis categóricas (fatores): “TC” e “TNC” ficaram com dois níveis, “EST” e “PLC” com três níveis, “THIS” com quatro níveis e “ID” com cinco níveis. A variável TM entrou no modelo como variável numérica.

Foram incluídas duas interações no modelo, tamanho do tumor/proporção de linfonodos comprometidos e tamanho do tumor/estadiamento.

A seguir apresentamos a codificação das variáveis realizada:

O fator TC (tipo de cirurgia a qual a paciente foi submetida), com 2 níveis: CQ (cirurgia conservadora de retirada do quadrante comprometido pelo tumor) e CMAST (cirurgia radical de retirada total da mama). A variável indicadora do fator TC foi distribuída como mostra a Tabela 1.

Tabela 1- Variável indicadora

	Variável indicadora
TC	TC1
CQ	0
CMAST	1

O modelo contendo o termo ℓ_h é ajustado pela inclusão da variável indicadora TC1. Assim, o efeito devido ao h-ésimo nível do fator TC será denotado por $\ell_h,\ h=1.$ Desta forma:

$\ell =$ o efeito diferencial do nível 1 (CMAST) com relação à categoria de referência (CQ).

O fator TNC (tipo de tratamento não-cirúrgico ao qual a paciente foi submetida), com 2 níveis: 0 (QHR ou QH ou RH ou H) e 1 (OUT ou QNEOR ou QR). A variável indicadora do fator TNC foi distribuída como mostra a Tabela 2.

Tabela 2- Variável indicadora do fator TNC

	Variável indicadora
TNC	TNC1
QRH ou QH ou RH ou H	0
OUT ou QNEOR ou QR	1

OBS: H- hormonioterapia, QH (quimio e hormonioterapia), QNEORH - quimioterapia neoadjvante e radio e hormonioterapia, QR - quimio e radioterapia, QRH - quimio-radio-hormonioterapia, RH - radio-hormonioterapia, OUT – neste grupo estão incluídos a quimioterapia ou radioterapia. Ressaltamos que onde não houver discriminado que o tratamento foi neoadjuvante (antes da cirurgia) é porque se trata de um tratamento adjuvante (após a cirurgia).

O modelo contendo o termo δ_p é ajustado pela inclusão da variável indicadora TNC1 Assim, o efeito devido ao p-ésimo nível do fator TNC será denotado por $\delta_p,\ p=1.$ Desta forma:

$\delta_1 =$ é o efeito diferencial do nível 1 (OUT+QNEORH+QR) com relação à categoria de referência (QRH + H + QH + RH).

O fator EST (estadiamento), com três níveis (0 + I + IIA; IIB; IIIA + IIIB). Com as seguintes variáveis indicadoras apresentadas na Tabela 3:

Tabela 3- Variáveis indicadoras do fator EST

	Variáveis indicadoras	
EST	E1	E2
0+ I + IIA	0	0
IIB	1	0
IIIA + IIIB	0	1

O estadiamento do tumor possui as seguintes classificações: O, I, IIA, IIB, IIIA e IIIB, da menor para a maior gravidade do tumor, respectivamente.

O modelo contendo o termo ν_r é ajustado pela inclusão das variáveis indicadoras E1 e E2. Assim, o efeito devido ao r-ésimo nível do fator EST será denotado por $\nu_r,\ r=1,2.$ Desta forma:

$\nu_1 =$ é o efeito diferencial do nível 1 (IIB) com relação à categoria de referência (0 + 1 + IIA).

$\nu_2 =$ é o efeito diferencial do nível 2 (IIIA + IIIB) com relação à categoria de referência (0 + 1 + IIA).

O fator PLC (proporção de linfonodos axilares comprometidos com metástase), com três níveis: 0 (nenhum linfonodo comprometido por metástase); >0 – 0,5 (proporção maior que zero e menor que cinquenta por cento) e >0,5 (mais que cinquenta por cento de linfonodos comprometidos por metástase). Com as seguintes variáveis indicadoras apresentadas na Tabela 4:

Tabela 4- Variável indicadora do fator PLC

	Variáveis indicadoras	
PLC	PLC1	PLC2
0	0	0
0—|0,5	1	0
>0,5	0	1

O modelo contendo o termo τ_k é ajustado pela inclusão das variáveis indicadoras PLC1 e PLC2. Assim, o efeito devido ao k-ésimo nível do fator PLC será denotado por $\tau_k,\ k = 1, 2.$ Desta forma:

$\tau_1 =$ é o efeito diferencial do nível 1 (>0 – 0,5) com relação à categoria de referência (0).

$\tau_2 =$ é o efeito diferencial do nível 2 (>0,5) com relação à categoria de referência (0).

O fator THIS (tipo histológico do tumor), com quatro níveis:LOB (lobular); CDINS (carcinoma *ductal in situ*); CDINF(carcinoma *ductal* infiltrante); OUT(outras formas). Com as seguintes variáveis indicadoras apresentadas na Tabela 5:

Tabela 5- Variáveis indicadoras do fator THIS

	Variáveis indicadoras		
THIS	H1	H2	H3
LOB	0	0	0
CDINS	1	0	0
CDINF	0	1	0
OUT	0	0	1

O modelo contendo o termo λ_n é ajustado pela inclusão das variáveis indicadoras H1, H2 e H3. Assim, o efeito devido ao n-ésimo nível do fator THIS será denotado por $\lambda_n,\ n = 1, 2, 3.$ Desta forma:

$\lambda_1 =$ é o efeito diferencial do nível 1 (CDINS) com relação ao nível de referência (LOB).

$\lambda_2 =$ é o efeito diferencial do nível 2 (CDINF) com relação à categoria de referência (LOB).

$\lambda_3 =$ é o efeito diferencial do nível 3 (OUT) com relação à categoria de referência (LOB).

O fator ID (faixa etária), com cinco níveis: <=35 (menor ou igual a trinta e cinco anos); 36 – 45 (trinta e seis a quarenta e cinco anos); 46 – 55 (quarenta e seis a cinquenta e cinco anos); 56 – 65 (cinquenta e seis a sessenta e cinco anos); >=66 (maior ou igual a sessenta e seis anos). Com as seguintes variáveis indicadoras apresentadas na Tabela 6.

Tabela 6- Variáveis indicadoras do fator ID

	Variáveis indicadoras			
ID	I1	I2	I3	I4
≤35	0	0	0	0
36—45	1	0	0	0
46—55	0	1	0	0
56—65	0	0	1	0
≥66	0	0	0	1

O modelo contendo o termo β_j é ajustado pela inclusão das variáveis indicadoras I1, I2, I3 e I4. Assim, o efeito devido ao j-ésimo nível do fator ID será denotado por $\beta_j,\ j = 1,2,3,4.$ Desta forma:

$\beta_1 =$ é o efeito diferencial do nível 1 (36 -- 45) com relação à categoria de referência (<=35).

$\beta_2 =$ é o efeito diferencial do nível 2 (46 -- 55) com relação à categoria de referência (<=35).

$\beta_3 =$ é o efeito diferencial do nível 3 (56 -- 65) com relação à categoria de referência (<=35).

$\beta_4 =$ é o efeito diferencial do nível 4 (>=66) com relação à categoria de referência (<=35).

A variável numérica TTUM (tamanho do tumor em cm), com coeficiente γ .

$\gamma =$ a alteração no logaritmo da razão de risco de recidiva quando acrescemos um cm no tamanho do tumor.

O modelo de regressão de Cox, com todas as covariáveis e interações, pode ser escrito da seguinte forma:

hi(t)= exp

$$\{\tau_1 PLC2.1+\tau_2 PLC2.2+\beta_1 I1+\beta_2 I2+\beta_3 I3+\beta_4 I4+\lambda_1 H1+\lambda_2 H2+\lambda_3 H3+\gamma TTUM+$$

$$\nu_1 E1+\nu_2 E2+\ell TC1+\delta_1 TNCRRR+\gamma\tau_1 TTUMPLC21+\gamma\tau_2 TTUMPLC22+\gamma\nu_1 TTUME1+$$
$$\gamma\nu_2 TTUME2\}\ h_0(t)$$

A função de risco base corresponde a uma paciente que possui PLC = 0 (nenhum linfonodo comprometido); EST = 0 + 1 + IIA; ID <= 35 anos; THIS = LOB (tipo histológico lobular); TC = CQ (quadrantectomia) e TNC = 0 (pacientes que se submeteram ao tratamento quimio, rádio e hormonioterapia, ou ao tratamento de radio e hormonioterapia, ou ao tratamento quimio e hormonioterapia ou só ao tratamento de hormonioterapia).

A interação, tamanho do tumor e proporção de linfonodos comprometidos (TTUM x PLC2), foi incluída no modelo pelos produtos TTUMPLC1 = (TTUM) x (PLC1) e TTUMPLC2 = (TTUM) x (PLC2). Possuem coeficientes $\gamma\tau_1$ e $\gamma\tau_2$, respectivamente. A interação, tamanho do tumor e estadiamento (TTUM x EST), foi incluída no modelo pelos produtos TTUME1= (TTUM) x (E1) e TTUME2= (TTUM) x (E2). Possuem coeficientes $\gamma\nu_1$ e $\gamma\nu_2$, respectivamente.

3.2. Aplicação

Ilustraremos, através de três variáveis, a etapa utilizada na categorização das variáveis. As funções de sobrevivências estimadas através do estimador Kaplan-Meier e a comparação de suas curvas de sobrevivência através do teste *log-rank*:

a) Tipo de Tratamento Cirúrgico

A covariável TCIR está categorizada da seguinte forma: no Grupo CQ estão as pacientes que se submeteram à cirurgia do tipo quadrantectomia e no Grupo CMAST estão aquelas que se submeteram à cirurgia do tipo mastectomia. A Figura 2 ilustra as curvas estimadas para os dois grupos.

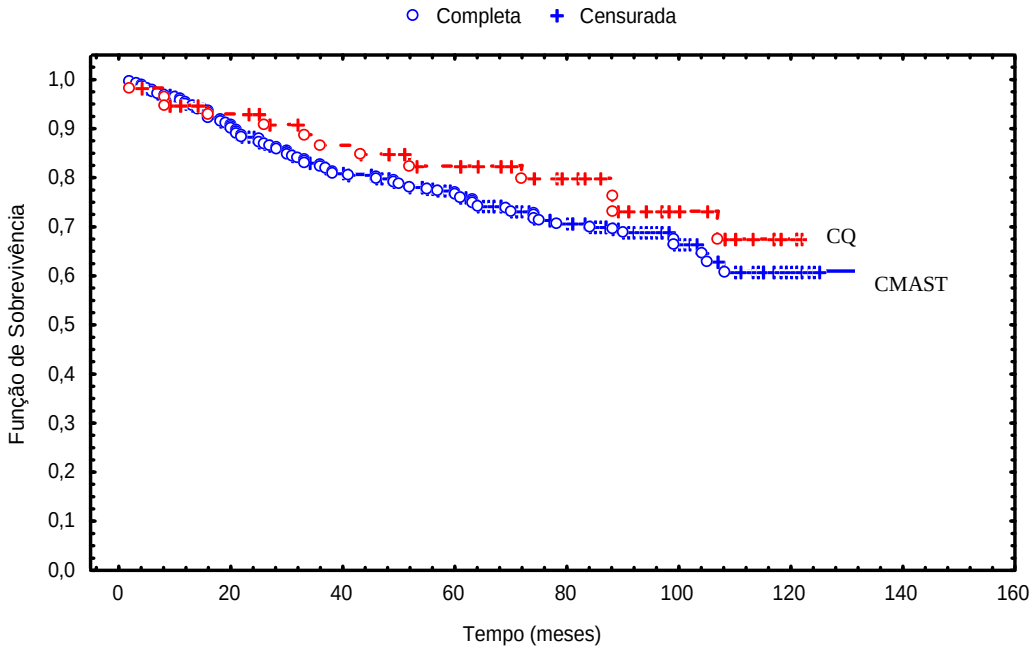


Figura 2: Curva de Sobrevivência (KM) segundo tipo de cirurgia (TCIR).

As hipóteses de interesse são:

H0 : não existe diferença no tempo livre da doença quando comparamos os dois grupos de pacientes.

H1: existe diferença.

Teste *log-rank* $\left(W_L = \frac{U_L^2}{V_L} \sim \chi_1^2 \right)$. WL= 0,8 com p-valor = 0,362

Conclusão: O gráfico aponta para um maior tempo livre da doença no Grupo CQ (pacientes submetidas a quadrantectomia). Porém o teste *log-rank* foi não significativo, ao nível de significância de 5%, ou seja, a este nível de significância não podemos rejeitar a hipótese nula *H0* de que não existe diferença no tempo livre da doença quando comparamos os dois Grupos de pacientes segundo tipo de cirurgia realizada.

b) Proporção de Linfonodos Axilares Comprometidos

A covariável PLC2 está categorizada da seguinte forma: no Grupo 0, estão as pacientes cuja proporção de linfonodos comprometidos foi igual a 0, ou seja, de todos os linfonodos ressecados nenhum apresentava metástase; no Grupo 1, estão aquelas cuja proporção de linfonodos comprometidos foi maior que 0 chegando até 50% ;e no Grupo 2, estão aquelas com mais de 50% dos linfonodos axilares comprometidos.

A Figura 3 ilustra as curvas estimadas para os três grupos.

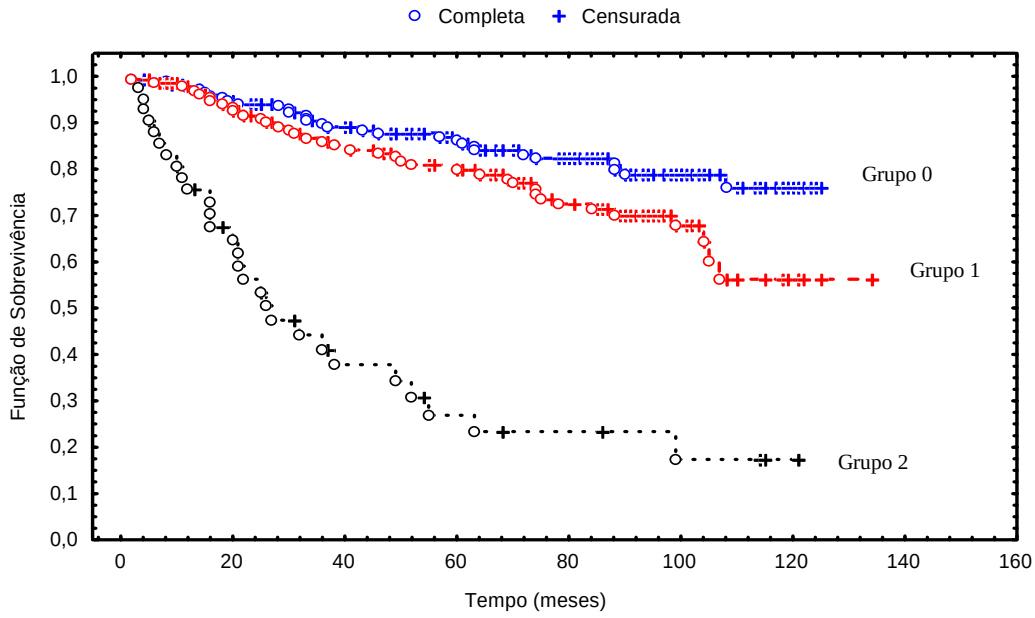


Figura 3: Curva de Sobrevivência (KM) segundo proporção de linfonodos comprometidos (PLC2).

As hipóteses de interesse são:

H0 : não existe diferença no tempo livre da doença quando comparamos os três grupos de pacientes.

H1: existe diferença.

Teste *log-rank* $\left(W_L = \frac{U_L^2}{V_L} \sim \chi_2^2 \right)$. WL= 85,3 com p-valor = 0,0000000

Conclusão: o gráfico mostra que o grupo 3 apresenta uma curva abaixo das curvas dos grupos 0 e 1, isto sugere que o tempo livre da doença é inferior no grupo das pessoas que apresentam mais de 50% dos linfonodos comprometidos. O gráfico aponta para um maior tempo livre da doença no grupo 0. O teste *log-rank* confirma uma diferença altamente significativa entre as curvas, ou seja, existe uma real diferença no tempo até a recidiva da doença quando comparamos os três grupos pertencentes às categorias da proporção de linfonodos comprometidos.

c) Tipo de Tratamento Não Cirúrgico

Apresentaremos a função de sobrevivência estimada através do estimador *Kaplan-Meier* e a comparação das curvas de sobrevivência através do teste *log-rank* para a covariável TRATNC (tipo de tratamento não cirúrgico).

A covariável TNCRRR foi categorizada da seguinte forma: no Grupo 0 ficaram as pacientes que se submeteram ao tratamento QRH, ou ao tratamento H, ou ao tratamento QH, ou ao tratamento RH, enquanto no Grupo 1 estão as pacientes que se submeteram somente à radioterapia, ou somente à quimioterapia, ou QR, ou ao tratamento QNEORH. A Figura 4 ilustra as curvas estimadas para os dois grupos.

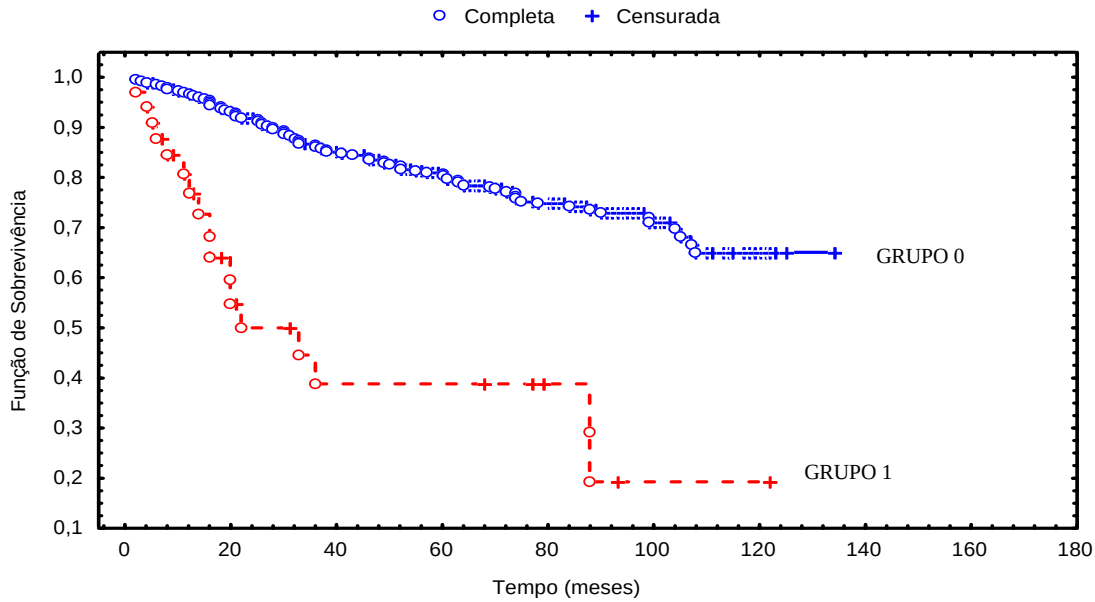


Figura 4: Curva de Sobrevivência (KM) segundo tipo de tratamento não cirúrgico (TNCRRR).

As hipóteses de interesse são:

H0 : não existe diferença no tempo livre da doença quando comparamos os dois grupos de pacientes.

H1: existe diferença.

Teste *log-rank* $\left(W_L = \frac{U_L^2}{V_L} \sim \chi_1^2 \right)$. WL= 41,2 com p-valor = 0,0000000

Conclusão: a análise gráfica indica que existe uma possível diferença entre as curvas de KM dos dois Grupos. O gráfico aponta para um maior tempo livre da doença no Grupo 0 (pacientes submetidas aos tratamentos QRH, H, QH, e RH). O teste *log-rank* confirma uma diferença altamente significativa entre as curvas, ou seja, existe ao nível de significância de 5%, uma diferença no tempo até a recidiva da doença quando comparamos os dois Grupos de pacientes segundo classificação do tratamento não cirúrgico.

4. Resultados e Discussão

Das 355 pacientes, a maioria, 259 (73%), não apresentou recidiva da doença (censuras). As demais, 96 pacientes representam falhas (recidiva da doença).

As 355 mulheres que compõem a amostra estudada apresentavam idade entre 24 a 91 anos com média de 56 anos, sendo o grupo etário de 46 a 55 anos o mais frequente.

A cirurgia radical (mastectomia) foi realizada na maioria das pacientes, somando um total de 297 mastectomias (83,7%). Considerando o número de linfonodos axilares retirados (1- 40), segundo o laudo anátomo-patológico, o número médio de linfonodos axilares comprometidos foi de três linfonodos. Das 355 pacientes, um total de 174,(49%), não apresentava nenhum linfonodo axilar comprometido com metástase.

Com relação ao tratamento não cirúrgico, a maioria das mulheres, um total de 198 (56%), recebeu o tratamento adjuvante (pós-cirúrgico) conjunto da quimio, rádio e hormonioterapia (QRH).

O tipo histológico mais comum do tumor foi o carcinoma *ducta*l infiltrante com 308 casos (87%).

Com relação ao estadiamento do tumor a classificação predominante foi IIA com 137 registros (39%). Em relação ao tamanho do tumor a média foi de 3,2 cm de diâmetro. Podemos constatar que o câncer de mama é diagnosticado em estágios tumorais mais avançados, pois os dois estágios iniciais (estadiamento 0 e 1) somados correspondem apenas a 46 casos (13%).

O tempo de acompanhamento foi entre 2 a 134 meses com média de 66 meses. O tempo mediano de acompanhamento foi de 70 meses, isto significa que pelo menos 50% das pacientes foram acompanhadas por 70 meses.

Foi selecionado o modelo mais adequado para o ajuste dos dados deste estudo. Permaneceram no modelo final as covariáveis que apresentaram uma estatística de verossimilhança com significância menor ou igual a 5%. O modelo de risco proporcional de Cox com as covariáveis selecionadas é dado por:

$$h_i(t) = \exp\left\{\tau_1 PLC2.1_i + \tau_2 PLC2.2_i + \delta_1 TNCRRR_i\right\} h_0(t), \quad i = 1,2,\ldots,n.$$

O ajuste dos dados com o modelo final proposto é apresentado na Tabela 7.

Tabela 7- Resultados do Ajuste do Modelo de Riscos Proporcionais de Cox.

Variáveis Indicadoras	Coeficientes (coef.)	Razão de Risco exp {coef.}	e.p.(coef.)	p-valor
PLC.1	0,486	1,630	0,244	**0,0470000**
PLC.2	1,927	6,870	0,274	**0,0000000**
TNC	1,270	3,560	0,283	**0,0000070**

O Teste da Razão de verossimilhança (ξRV) foi igual a 67,7 com 3 graus de liberdade p-valor <0,001> e amostra de 355 pacientes.

Os resultados disponíveis na Tabela 7 mostram o papel das covariáveis PLC (proporção de linfonodos comprometidos por metástase) e TNC (tipo de tratamento não-cirúrgico) na predição do tempo de recidiva, ou seja, na predição do tempo livre do câncer de mama em pacientes submetidas ao tratamento cirúrgico para retirada do tumor mamário. O modelo ajustado para PLC2 (PLC1+ PLC2) e para TNC conduz a uma redução de 67,7 no valor do

-2log L (teste da razão de verossimilhança), a qual é altamente significante quando comparado com o percentil da distribuição qui-quadrado com 3 graus de liberdade (ξRV = 67,7; p-valor = 0,000000000). O que parece correto afirmar que as pacientes com proporção dos linfonodos comprometidos superior a 50% têm um risco de recidiva da doença aproximadamente sete vezes maior quando comparadas com aquelas cuja proporção de linfonodos comprometidos foi igual a zero (p-valor = 0,0000000). As pacientes com proporção dos linfonodos comprometidos superior a zero atingindo até 50% têm um risco de recidiva da doença aproximadamente 1,6 vezes maior quando comparadas com aquelas cuja proporção de linfonodos comprometidos foi igual a zero (p-valor = 0,047). As pacientes que se submeteram aos tratamentos de radioterapia, ou quimioterapia, quimioneoadjuvante mais rádio e hormonioterapia, ou quimio mais radioterapia têm um risco de recidiva da doença aproximadamente 3,5 vezes maior quando comparadas com as pacientes do Grupo 0 (aquelas que se submeteram aos tratamentos quimio mais rádio mais hormonioterapia, hormonioterapia, quimio mais hormonioterapia, rádio mais hormonioterapia (p-valor = 0,000007).

5. Conclusões

Os resultados das análises obtidas neste trabalho confirmam que o comprometimento dos linfonodos é um fator de extrema importância na predição da recidiva do câncer de mama. As análises indicam que as pacientes com proporção de linfonodos comprometidos superior a 50% têm um risco de recidiva da doença aproximadamente sete vezes maior quando comparadas com aquelas sem linfonodos comprometidos. As pacientes com proporção de linfonodos comprometidos no intervalo maior que zero até 50% têm um risco de recidiva da doença aproximadamente 1,6 vezes maior quando comparadas com aquelas cuja proporção de linfonodos comprometidos foi igual a zero. As pacientes que se submeteram aos tratamentos radioterapia, ou quimioterapia, quimioneoadjuvante mais rádio e hormonioterapia, ou quimio mais radioterapia têm um risco de recidiva da doença aproximadamente 3,5 vezes maior quando comparadas com as pacientes do Grupo 0 (aquelas que se submeteram aos tratamentos quimio mais rádio mais hormonioterapia, hormonioterapia, quimio mais hormonioterapia, rádio mais hormonioterapia). Isto indica que a hormonioterapia parece ser o elemento diferenciador que contribui no aumento do tempo livre de recidiva do câncer de mama, quando associado a tratamentos de radioterapia ou quimioterapia adjuvante. Entretanto, a hormonioterapia não parece ser capaz de aumentar o tempo livre da doença nas pacientes submetidas à quimioterapia neoadjuvante, de forma que estas pacientes atinjam um tempo livre da doença igual as pacientes do Grupo 0. Não existe diferença no tempo de recidiva quando comparamos pacientes que se submeteram à mastectomia daquelas que se submeteram à quadrantecomia.

Os dados refletem uma situação preocupante com relação ao diagnóstico precoce do tumor de mama. A maioria dos casos é diagnosticada nos estádios mais avançados (II,III e IV). Neste estudo, constatamos que, das 485 pacientes consideradas na população total atendida no Hospital Dr. Luiz Antônio, 123 (25%) se encontravam nos estádios III ou IV, 281 (58%) nos estádios IIA e IIB e apenas 61 (13%) nos estádios iniciais 0 ou 1. São necessárias medidas urgentes para assegurar que programas de rastreamento sejam acessíveis a toda população.

A identificação de fatores que possam melhor conduzir o tratamento no sentido de prevenir a recidiva do câncer de mama é de extrema importância na redução da mortalidade causada por esta doença.

Referências bibliográficas

ANDERSEN, P. K. Testing goodness of fit of Cox' regression and life model. Biometrika, v.38, 1982. p.67-77.

ALBERG, A. J., VISVANATHAN, K., HELZLSOUER, K. J. Epidemiology, prevention, and early detection of breast cancer.1998.

BERGMANN, A. Prevalência do linfedema subseqüente a tratamento cirúrgico para câncer de mama no Rio de Janeiro. (Mestrado). Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública, 2000. 142p.

BRESLOW, N. E., DAY, N. E. Statistical methods in cancer research; The analysis of case-control studies. v.1, Lyon: IARC Scientific Publications, 1980. 350p.

BUSTAMANTE-TEIXEIRA, M.T., FAERSTEIN, E., LARORRE, M.R.D. Técnicas de Análise de Sobrevivência. Cadernos de Saúde Pública. 2001. 34p.

COLLETT, D. Modelling Survival Data in Medical Research. 1.ed. London: Chapman & Hall, 1994. 347p.

COLOSIMO, E. A. Análise de Sobrevivência Aplicada. In: 46ª REUNIÃO ANUAL DA REGIÃO DA SOCIEDADE INTERNACIONAL DE BIOMETRIA (RBRAS) e 9º SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO Agronômica (SEAGRO). Piracicaba, 2001. 145p.

COX, D.R. Regression models and life-tables. Journal of Royal Statistical Society B, v.34, p.187-220, march, 1972.

COX, D. R. Partial likehood. Biometrika, v.62, p. 269-276, march, 1975.

FLEMING, T.R., HARRINGTON, D.P. Couting Processes e Survival Analysis. Canadá, 1991.429p.

GALEB JR., N. A., GARRIDO, M. M., DE LUCA, L.A., OSÓRIO, C.A.B. T. COSTA, R. L. R., GAMAEIRO, P. L., GOÉS, J. C. S. Estudos das técnicas para pesquisa do linfonodo sentinela no câncer de mama. Revista Brasileira de Mastologia, v.10, n.3 p.107-14, 2000.

_____GILL R.; SCHUMACHER M. A simple test of the proportional hazards assumption

_____Biometrika 1987 74: 289-300; doi:10.1093/biomet/74.2.289

INSTITUTO NACIONAL DO CÂNCER (INCA/MS). Acesso www.inca.gov.br. 2005.

INSTITUTO NACIONAL DO CÂNCER (INCA/MS). Câncer no Brasil. Dados dos Registros de Base Populacional. Brasília: Ministério da Saúde, 2003.

KALBFLEISCH, J. D., PRENTICE, R. L. Marginal likelihoods based on Cox's regression and life model. Biometrika, v.60, p.267-78, 1973.

MALLER, A. R., ZHOU, X. Survival Analysis with Long-Term Survivors. Inglaterra, 1996. 278p.

Abstract

Statistical techniques of survival analysis treat data relating a time until the occurrence of the phenomenon. This article uses the survival analysis model, the Cox Regression Model (Cox,1972). The supposition of proportional risks was considered in the group of analyzed data. The breast cancer is a serious problem of Public Health in Brazil. It causes a high taxe of death among women. Our principal aim was to check among registred informations, that is, to choose among some explaining variables explanations whose could be so significant for the time in what the patient, with breast cancer, could stay free of the returning (falling back) of the disease.

Keywords: Survival analysis, Cox regression model, breast cancer.

Análise e Implementação de Redes Neurais Generalizadas

Guilherme Guimarães Moreira[8]
Marcelo Azevedo Costa[9]

Resumo

Este artigo propõe os modelos de redes neurais generalizadas. Estes modelos agregam a estrutura de verossimilhança dos modelos lineares generalizados e a flexibilidade das redes neurais artificiais na modelagem de interações não-lineares e não-aditivas entre as variáveis preditoras e a variável resposta. O treinamento é realizado segundo o método iterativo do gradiente descendente, que procura minimizar a função desvio do modelo. O critério de qualidade do modelo é obtido via validação cruzada. Os resultados preliminares sugerem que as redes neurais generalizadas podem apresentar resultados de previsão melhores ou equivalentes aos modelos lineares generalizados.

[8]Instituto Brasileiro de Geografia e Estatística.
[9] Universidade Federal de Minas Gerais.

1. Introdução

As Redes Neurais Artificiais (RNAs) têm recebido grande atenção por parte de pesquisadores de diversas áreas, sendo utilizadas nos mais diversos problemas de modelagem de bases de dados. Na estatística, as RNAs são cada vez mais utilizadas em problemas de classificação e predição, em virtude da sua capacidade de representação interna de informação caracterizada pelo paralelismo inerente à sua arquitetura, possibilitando desempenho de predição superior aos modelos convencionais. Como consequência, as RNAs conseguem modelar efeitos não-lineares e não-aditivos das covariáveis em relação à variável resposta, além de outras relações existentes entre as próprias covariáveis, as quais não foram relacionadas *a priori* pelo pesquisador na confecção de um modelo estatístico convencional.

Problemas de modelagem estatística, tais como: classificar um determinado indivíduo em um grupo, ou predizer a sua resposta a um determinado tratamento com base em algumas de suas características podem ser modelados segundo a teoria dos Modelos Lineares Generalizados (NELDER e WEDDERBURN, 1972). Estes modelos permitem realizar a modelagem estatística de dados multivariados, associando-se à variável resposta uma distribuição definida na família exponencial. Entretanto, em algumas situações, a escolha de um modelo pode ser um problema complexo, devido à falta de informação sobre a variável de interesse e sua correlação com as demais variáveis preditoras e/ou à existência de muitos fatores não-lineares e não-aditivos a serem estimados (BIGANZOLI *et al*., 1998). Em tais situações, pode ser mais apropriado considerar modelos flexíveis que sejam capazes de proporcionar uma resposta coerente de predição, seja para um problema de regressão ou classificação, representando internamente as várias correlações existentes e as desconhecidas ao pesquisador. Neste contexto, as RNAs podem ser consideradas como modelos flexíveis apropriados para a resolução de problemas multivariados não-lineares.

Segundo Biganzolli *et al.* (1998), as RNAs podem ser consideradas como uma generalização não-linear dos MLGs. Em função de sua característica computacional, a avaliação do potencial da resposta produzida pelas RNAs para predição e/ou classificação deve ser baseada na comparação empírica das redes com as obtidas por outros métodos estatísticos aplicados a dados reais.

Em particular, as redes neurais artificiais têm sido exploradas na análise de dados médicos (DYBOWSKI, 2001), Biganzolli *et al*. (1998) e Biganzoli *et al*. (2006) apresentam as RNAs para modelos de análise de sobrevivência, utilizando para isso uma adaptação do modelo de sobrevivência de Cox (1984), onde o preditor linear do modelo é substituído por uma rede do tipo MLP (*Multi-Layer-Perceptron*). Outras variações e aplicações de RNAs para

modelos de sobrevivência podem ser encontrados em Ambrogi *et al.* (2007), Zhang & Zhang (2008) e Lisboa *et al.* (2008).

Neste trabalho, o modelo de RNA é estendido para os modelos da família exponencial, definindo-se então as Redes Neurais Generalizadas. As RNGs se distinguem dos modelos de RNAs convencionais pela incorporação da função de verossimilhança da família exponencial à estrutura não-linear da RNA. Este modelo é, então, mais amplo do que a proposta apresentada por Biganzolli *et al.* (1998). Na sequência, a partir da formulação deste modelo, é avaliada a sua capacidade de produzir respostas precisas quando comparadas àquelas fornecidas por modelos convencionais.

O trabalho encontra-se organizado da seguinte forma: a seção 2 apresenta uma revisão das redes MLPs e do algoritmo *backprogapation*. Na seção 3, as RNGs são descritas e o algoritmo *backpropagation* é definido para o novo modelo. A seção 4 descreve a metodologia de comparação das RNGs com os MLGs e apresenta dois resultados comparativos para duas bases de dados reais. A conclusão do trabalho é apresentada na seção 5.

2. Redes Neurais Artificiais

As RNAs são sistemas paralelos distribuídos compostos por unidades de processamento simples (nodos ou neurônios) que calculam determinadas funções matemáticas. Tais unidades estão dispostas em uma ou mais camadas e interligadas. Na maioria dos modelos, estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede (BRAGA, CARVALHO e LUDERMIR, 2000).

A solução de problemas via RNAs é bastante atrativa, já que a forma como estes são representados internamente pela rede e o paralelismo natural inerente à sua arquitetura criam a possibilidade de um desempenho superior ao dos modelos convencionais, pois ela consegue estimar tanto fatores lineares quanto não-lineares, o que aumenta o seu poder de predição e classificação frente a outros modelos.

Para exemplificarmos, reproduziremos o exemplo de Dugas *et al.,* (2003). Suponha duas variáveis de entrada, x_1 e x_2. Nos modelos de regressão linear clássicos, uma forma de tornar o modelo mais flexível consiste em incluir combinações não-lineares fixas entre os

regressores, tais como: x_1^2, x_2^2, $x_1 x_2$, $x_1^2 x_2$, ... Entretanto, esta aproximação adiciona exponencialmente muitos termos à regressão à medida que a ordem do polinômio aumenta.

Em contrapartida, considere um único neurônio existente na camada escondida de uma RNA, conectada a duas entradas. Os parâmetros da rede ajustáveis são α_0, α_1 e α_2. Uma função típica para este neurônio é dada pela função *tangente hiperbólica*:

$$\tanh(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2).$$

executando uma expansão da série de Taylor em *y* para $\tanh(\alpha_0 + y)$, onde $y = \alpha_1 x_1 + \alpha_2 x_2$, e considerando $\beta = \tanh(\alpha_0)$, tem-se:

$$\tanh(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2) = \beta + (1-\beta^2)(\alpha_1 x_1 + \alpha_2 x_2) + (-\beta + \beta^3)(\alpha_1 x_1 + \alpha_2 x_2)^2 + \left(-\frac{1}{3} + \frac{4\beta^2}{3} - \beta^4\right)(\alpha_1 x_1 + \alpha_2 x_2)^3 + \left(\frac{2\beta}{3} - \frac{5\beta^2}{3} + \beta^5\right)(\alpha_1 x_1 + \alpha_2 x_2)^4 + O(\alpha_1 x_1 + \alpha_2 x_2)^5$$

Apesar de o número de termos ser infinito, a função não-linear computada por este único neurônio inclui todas as representações das variáveis da entrada, porém elas não podem ser todas controladas independentemente. Os termos dependem apenas dos coeficientes α_0, α_1 e α_2. Dessa forma, adicionar mais neurônios à rede RNA aumenta a flexibilidade da função computada pela rede, uma vez que cada neurônio conectado permite que o modelo capture tantos relacionamentos não-lineares entre as variáveis quanto o número de neurônios disponíveis.

O procedimento usual de solução de problemas em RNAs passa inicialmente por uma fase de especificação de topologia: define-se o número de neurônios, a disposição dos mesmos e o algoritmo de aprendizagem, com o qual um conjunto de padrões ou uma amostra é apresentado para a rede que, por sua vez, extrai as características necessárias para representar a informação fornecida e, posteriormente, gerar respostas coerentes para o problema.

A esta capacidade de aprender através de exemplos e de generalizar a informação aprendida é, sem dúvida, o principal atrativo à solução de problemas via RNAs. A generalização está associada à capacidade de a rede aprender através de um conjunto reduzido de exemplos e, posteriormente, inferir resultados para dados desconhecidos. Não obstante, as redes também são capazes de atuar como mapeadores universais de funções multivariáveis, com um custo computacional que cresce linearmente de acordo com o número de variáveis.

2.1 Redes Neurais Artificiais do tipo *Multi Layer Perceptron*

As redes *Multi Layer Perceptron* – MLP, são caracterizadas por uma camada de entrada, uma de saída e uma ou mais camadas intermediárias ou ocultas. Segundo Cybenko (1989), uma rede com uma camada intermediária é capaz de implementar qualquer função contínua, enquanto a utilização de duas camadas intermediárias permite a aproximação de qualquer função. As redes MLPs têm sido aplicadas com sucesso para resolver diversos problemas complexos através de seu treinamento supervisionado, o qual utiliza um algoritmo muito popular conhecido como algoritmo *backpropagation*.

Neste artigo, nos restringimos a uma rede MLP com apenas uma camada intermediária, Figura (2.1). Nessa rede todos os neurônios de entrada estão ligados a todos os neurônios da camada intermediária, e estes, por sua vez, estão todos conectados aos da camada de saída. Nos neurônios da camada intermediária, uma transformação não-linear é realizada sobre a soma ponderada das entradas. Considere o i-ésimo vetor de entradas $x_i = [x_{i1}, \cdots, x_{iP}]^T$, onde *P* é o número de covariáveis do modelo. Considere também uma matriz W1 de pesos, onde $w1_{ph}$ é o peso associado à p-ésima covariável e ao h-ésimo neurônio da camada intermediária. Considere também $W2 = [w_1, \cdots, w_H]^T$ e $\beta 1 = [\beta_1, \cdots, \beta_H]^T$, vetores de pesos e constantes associados a cada neurônio da camada intermediária, respectivamente, sendo *H* o número de neurônios da camada intermediária. E por fim, $\beta 2$, uma constante. A entrada do h-ésimo neurônio da camada intermediária é dada pela projeção linear $W1^T x_i$ enquanto a sua saída é $\varphi(W1^T x_i)$, onde $\varphi(\cdot)$ é a 'função de ativação'. A função de ativação mais comumente utilizada é a função logística $\varphi(z) = \frac{1}{1+e^{-z}}$. Existem outras funções de ativação como, por exemplo, a tangente hiperbólica, a qual será a função de ativação utilizada neste artigo. A saída da rede é uma soma ponderada (por *W* 2) da saída dos neurônios da camada intermediária mais a constante $\beta 2$, já que nossa rede possui saída linear. Portanto, podemos representar uma rede neural com função de ativação sigmoidal (*tanh*) e saída linear como:

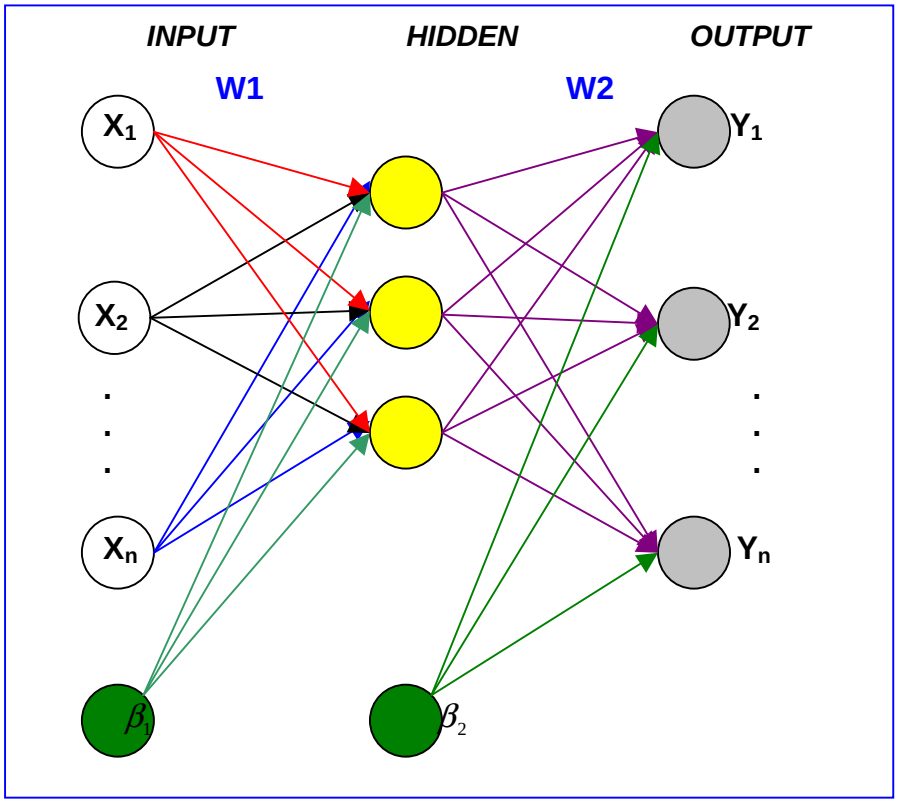
$$y_i = g(x_i;\theta) = \sum_{h=1}^{H} w2_h \tanh\left(\sum_{p=1}^{P} w1_{hp}\, x_{ip} + \beta 1_h\right) + \beta 2 \qquad (2.1)$$

onde,

θ é o vetor de parâmetros da rede $[\mathrm{W1}, \mathrm{W2}, \beta 1, \beta 2]^{\mathrm{T}}$

O número de parâmetros em uma rede como esta é *m* = HP + 2H + 1. Podemos definir **y** = g(x,θ) = $[g(x_1, \theta), ..., g(x_n, \theta)]^T$ como sendo o vetor composto por todas as *n* saídas da rede.

Figura 2.1 - Rede Neural Artificial do tipo MLP



Como descrito anteriormente, a rede MLP pode ser treinada pelo algoritmo *backpropagation*, o qual está descrito na seção seguinte.

2.1.1 Algoritmo *Backpropagation*

Existem diversos algoritmos para o treinamento de redes do tipo MLP (BRAGA *et al.*, 2006; COSTA e BRAGA, 2006). Dentre estes, o mais conhecido é o *backpropagation* (RUMELHART, HILTON e WILLIAMS, 1986). Este algoritmo é baseado na regra delta (HAYKIN, 2001), sendo por isto também denominado regra delta generalizada. O

algoritmo *backpropagation* propõe uma forma de definir o erro dos nodos das camadas intermediárias, possibilitando o ajuste de seus pesos através do método do gradiente descendente.

A função de custo a ser minimizada é a soma dos quadrados dos erros, descrita pela Equação (2.2).

$$J = \frac{\sum_{i=1}^{n} (y_i - d_i)^2}{2}, \qquad (2.2)$$

onde,

d_i é a i-ésima saída desejada e y_i é a i-ésima saída da RNA.

Considerando η, como sendo a taxa de aprendizado e sendo a saída da rede *y* dada pela Equação (2.1), temos então que a equação geral de ajuste dos pesos (parâmetros) da rede é dada por:

$$w_{(k+1)} = w_{(k)} - \eta \nabla J_{(k)}, \qquad (2.3)$$

onde, *k* é a iteração e ∇J o gradiente da função de custo.

Utilizando a regra da cadeia, é dado por:

$$\nabla J_{(k)} = \frac{\partial J}{\partial w_{(k)}} = \frac{\partial J}{\partial y_{(k)}} \cdot \frac{\partial y_{(k)}}{\partial w_{(k)}}. \qquad (2.4)$$

Logo, as quatro equações de ajuste dos pesos podem ser definidas para cada vetor de pesos associados às camadas escondidas e de saída da rede, como sendo:

$$\begin{aligned}
w1_{hp(k+1)} &= w1_{hp(k+1)} - \eta \frac{\partial J_{(k)}}{w1_{hp(k+1)}} \\
w2_{h(k+1)} &= w2_{h(k)} - \eta \frac{\partial J_{(k)}}{w2_{h(k)}} \\
\beta 1_{h(k+1)} &= \beta 1_{h(k)} - \eta \frac{\partial J_{(k)}}{\beta 1_{h(k)}} \\
\beta 2_{(k+1)} &= \beta 2_{(k)} - \eta \frac{\partial J_{(k)}}{\beta 2_{(k)}}
\end{aligned} \qquad (2.5)$$

onde

$$\begin{aligned}
\frac{\partial J_{(k)}}{w1_{hp(k)}} &= \frac{\partial J_{(k)}}{\partial y_{(k)}} \cdot \frac{\partial y_{(k)}}{w1_{hp(k)}} \\
\frac{\partial J_{(k)}}{w2_{h(k)}} &= \frac{\partial J_{(k)}}{\partial y_{(k)}} \cdot \frac{\partial y_{(k)}}{w2_{h(k)}} \\
\frac{\partial J_{(k)}}{\beta 1_{h(k)}} &= \frac{\partial J_{(k)}}{\partial y_{(k)}} \cdot \frac{\partial y_{(k)}}{\beta 1_{h(k)}} \\
\frac{\partial J_{(k)}}{\beta 2_{(k)}} &= \frac{\partial J_{(k)}}{\partial y_{(k)}} \cdot \frac{\partial y_{(k)}}{\beta 2_{(k)}}
\end{aligned} \qquad (2.6)$$

e,

$$\begin{aligned}
&\frac{\partial J_{(k)}}{\partial y_{(k)}} = d_{(k)} - y_{(k)} = e_{(k)} \\
&\frac{\partial y_{(k)}}{w1_{hp(k)}} = w2_{h(k)} \cdot \sec h^2\left(w1_{hp(k)} x_p + \beta 1_{h(k)}\right) \cdot x_p \\
&\frac{\partial y_{(k)}}{w2_{h(k)}} = \tanh\left(\sum_{p=1}^{P} w1_{hp(k)} x_p + \beta 1_{h(k)}\right) \\
&\frac{\partial y_{(k)}}{\beta 1_{h(k)}} = w2_{h(k)} . \sec h^2\left(\sum_{p=1}^{P}\left(w1_{hp(k)} x_p + \beta 1_{h(k)}\right)\right) \\
&\frac{\partial y_{(k)}}{\beta 2_{(k)}} = 1
\end{aligned} \qquad . \qquad (2.7)$$

Denotando de forma matricial temos:

$$\begin{aligned}
&\frac{\partial J_{(k)}}{W1_{(k)}} = -\left(W2_{(k)}^{T} \cdot e\right) \otimes \left(\sec h^2\left(W1 \cdot X + \beta 1_{(k)} \cdot 1_n\right)\right) \cdot X^T \\
&\frac{\partial J_{(k)}}{W2_{(k)}} = -e \cdot \left(\tanh\left(W1 \cdot X + \beta 1_{(k)} 1_n\right)\right)^T \\
&\frac{\partial J_{(k)}}{\beta 1_{(k)}} = -\left(W2_{(k)}^{T} \cdot e\right) \otimes \left(\sec h^2\left(W1 \cdot X + \beta 1_{(k)} \cdot 1_n\right)\right) \\
&\frac{\partial J_{(k)}}{\beta 2_{(k)}} = -e
\end{aligned} \qquad . \qquad (2.8)$$

onde $e_{(k)} = \left[\left(d_1 - y_{1(k)}\right), \cdots, \left(d_n - y_{n(k)}\right)\right]$, $1_n = [1, \cdots, 1]$ de dimensão n, $X = [x_1, \cdots, x_n]$ é uma matriz $p \times n$, tal que p é o número de co-variáveis e n o número de dados da amostra e '$\otimes$' simboliza o produto matricial termo a termo (se $A \otimes B = C$, então $c_{ij} = a_{ij} \times b_{ij}$).

As desvantagens do algoritmo *backpropagation* consistem em sua baixa velocidade de

convergência e na sua limitação ao se deparar com mínimos locais, já que se trata de um algoritmo que depende somente do gradiente local, ou seja, caso haja um ponto de mínimo local nas proximidades ele ficará preso.

Uma forma de se melhorar a resposta da rede utilizando o *backpropagation* é treinando diversas redes com pesos inicializados aleatoriamente, o que leva a rede a ter uma maior chance de obter uma solução que esteja próxima do mínimo global, porém isto implica em um grande custo computacional e não é possível garantir uma boa solução.

Vários métodos foram desenvolvidos a partir deste algoritmo visando evitar a convergência para regiões de mínimo local. O *RProp* (RIEDMILLER e BRAUN, 1993) utiliza o sinal do gradiente, e não o seu valor, para realizar a correção dos pesos. Já o *QuickProp* (FAHLMAN,1988) aproxima a superfície do erro por uma parábola em função dos pesos. Tal ajuste é realizado de forma que o erro mínimo da parábola seja alcançado. Porém, nem sempre esta superfície pode ser modelada por uma parábola. Os métodos de *Taxa Adaptativa* (SILVA e ALMEIDA,1990 e TOLLENAERE, 1990) utilizam técnicas para o ajuste da taxa de aprendizado. Porém, nenhum destes métodos é capaz de garantir uma convergência ao mínimo global. Braga *et al*. (2006) e Costa e Braga (2006) apresentam uma comparação de algoritmos ótimos para uma variedade de bases de dados. Os resultados mostram diferenças de velocidade de convergência e desempenho entre os algoritmos. Contudo, o algoritmo *backpropagation* com taxa de aprendizado reduzida e validação cruzada é capaz de gerar resultados ótimos, apesar da baixa velocidade de convergência.

3. Redes Neurais Generalizadas

Conforme descrito, a solução de problemas via RNAs é bem atrativa, já que a mesma consegue extrair dos dados, automaticamente, características sobre quais co-variáveis e suas interações são mais importantes sem que para isso seja necessário inferir previamente sobre essas relações. A idéia dos modelos de Redes Neurais Generalizadas - RNGs é agregar a informação da função de verossimilhança à modelagem não-paramétrica computacional das RNAs.

Seja a log-verossimilhança de um Modelo Linear Generalizado:

$$l(d,\theta)=\sum_{i=1}^{n}\frac{[d_i\theta_i-b(\theta_i)]}{a(\phi)}+c(d_i,\phi) \qquad (3.1)$$

onde, θ é o parâmetro referente à média, $\mu=b'(\theta)$, e ϕ é o parâmetro referente à variabilidade. Uma RNG é obtida a partir da associação do parâmetro canônico, θ, à saída de uma rede neural com função de ativação sigmoidal e saída linear, podendo ser expressa na forma:

$$\theta_i=\sum_{h=1}^{H}w2_h.\tanh\left(\sum_{p=1}^{P}w1_{hp}x_p+\beta1_h\right)+\beta2 \qquad (3.2)$$

ou, na forma matricial,

$$\theta=W2.\tanh(W1.X+\beta1\cdot1_n)+\beta2 \qquad (3.3)$$

Neste caso, pode-se utilizar o algoritmo de treinamento *backpropagation* para realizar a atualização dos pesos via método do gradiente descendente. Este método terá como objetivo maximizar a função de verossimilhança ou minimizar a função desvio com base em um conjunto de treinamento $T=\{(x_i,d_i)\}_{i=1}^{n}$. Dada uma certa condição inicial, $w_{(0)}$, para os pesos da rede, deseja-se obter a direção do ajuste a ser aplicado no vetor de pesos de forma a encontrarmos a direção para a solução, a qual maximiza a verossimilhança. A direção do ajuste no passo *k* pode ser obtida pelo gradiente da função de custo no ponto $w_{(k)}$.

A fim de maximizar o logaritmo da verossimilhança, o ajuste neste caso é realizado na mesma direção do gradiente no ponto $w_{(k)}$, ou seja, $w_{(k+1)}=w_{(k)}+\eta\,\Delta w_{(k)}$ onde: $\Delta w_{(k)}=\nabla l(d,\theta).$,

As equações de ajuste do vetor de parâmetros via método do gradiente, considerando um modelo neural como saída do parâmetro canônico, podem ser obtidas na forma:

$$
\begin{aligned}
w1_{hp(k+1)} &= w1_{hp(k)} + \eta r \frac{\partial l\left(d, \theta_{(k)}\right)}{w1_{hp(k)}} \\
w2_{h(k+1)} &= w2_{h(k)} + \eta r \frac{\partial l\left(d, \theta_{(k)}\right)}{w2_{h(k)}} \\
\beta 1_{h(k+1)} &= \beta 1_{h(k)} + \eta r \frac{\partial l\left(d, \theta_{(k)}\right)}{\beta 1_{h(k)}} \\
\beta 2_{(k+1)} &= \beta 2_{(k)} + \eta r \frac{\partial l\left(d, \theta_{(k)}\right)}{\beta 2_{(k)}}
\end{aligned}
\tag{3.4}
$$

onde, *r* é uma variável aleatória uniformemente distribuída no intervalo $[0,1]$, a qual incorpora um comportamento aleatório à taxa de aprendizado.

As derivadas da log-verossimilhança em relação aos parâmetros são definidas da seguinte forma:

$$
\begin{aligned}
\frac{\partial l\left(d, \theta_{(k)}\right)}{w1_{hp(k)}} &= \frac{\partial l\left(d, \theta_{(k)}\right)}{\partial \theta_{(k)}} \cdot \frac{\partial \theta_{(k)}}{w1_{hp(k)}} \\
\frac{\partial l\left(d, \theta_{(k)}\right)}{w2_{h(k)}} &= \frac{\partial l\left(d, \theta_{(k)}\right)}{\partial \theta_{(k)}} \cdot \frac{\partial \theta_{(k)}}{w2_{h(k)}} \\
\frac{\partial l\left(d, \theta_{(k)}\right)}{\beta 1_{h(k)}} &= \frac{\partial l\left(d, \theta_{(k)}\right)}{\partial \theta_{(k)}} \cdot \frac{\partial \theta_{(k)}}{\beta 1_{h(k)}} \\
\frac{\partial l\left(d, \theta_{(k)}\right)}{\beta 2_{(k)}} &= \frac{\partial l\left(d, \theta_{(k)}\right)}{\partial \theta_{(k)}} \cdot \frac{\partial \theta_{(k)}}{\beta 2_{(k)}}
\end{aligned}
\tag{3.5}
$$

onde $\frac{\partial l(d,\theta)}{\partial \theta}$ é a derivada da função custo em relação ao parâmetro canônico, θ, ou seja, é a derivada da função de log-verossimilhança em relação à saída da rede, e nos informa a contribuição da verossimilhança para o modelo de RNGs. A derivada $\frac{\partial l(d,\theta)}{\partial \theta}$ possui uma forma geral conhecida para os MLGs com ligação canônica dada por: $\frac{\partial l(d,\theta)}{\partial \theta} = d - \mu$, onde $d = [d_1, \cdots, d_n]^T$ com d_i sendo o valor desejado para i-ésima saída da rede e $\mu = [\mu_1, \cdots, \mu_n]^T$ com $\mu_i = b'(\theta_i)$ sendo o termo definido unicamente pela forma paramétrica associada à variável resposta.

O segundo termo das equações de ajuste dos pesos representa a derivada parcial da saída linear da rede MLP em relação aos vetores de pesos (ou parâmetros) definidas pelas Equações (2.7) e expresso em sua forma matricial pelas Equações (2.8).

Portanto, a forma do ajuste dos pesos para uma RNG é definida pelas seguintes expressões:

$$
\begin{aligned}
\Delta w1_{hp(k)} &= \left(d_{(k)} - y_{(k)}\right) \cdot w2_{h(k)} \cdot \sec h^2\left(w1_{hp(k)} x_p + \beta 1_{h(k)}\right) \cdot x_p \\
\Delta \beta 1_{h(k)} &= \left(d_{(k)} - y_{(k)}\right) \cdot w2_{h(k)} . \sec h^2\left(\sum_{p=1}^{P}\left(w1_{hp(k)} x_p + \beta 1_{h(k)}\right)\right) \\
\Delta w2_{h(k)} &= \left(d_{(k)} - y_{(k)}\right) \cdot \tanh\left(\sum_{p=1}^{P} w1_{hp(k)} x_p + \beta 1_{h(k)}\right) \\
\Delta \beta 2_{(k)} &= \left(d_{(k)} - y_{(k)}\right)
\end{aligned}
\qquad (3.6)
$$

onde

sech^2 é o quadrado da função secante hiperbólica ou a derivada da função tangente hiperbólica e

$d_{(k)} - y_{(k)}$ é a medida do erro entre a resposta desejada e a resposta predita pelo modelo.

O algoritmo *backpropagation* para a aplicação da RNG é dado a seguir:

Algoritmo de uma RNG

1. Inicialize os pesos: $W1$, $W2$, $\beta 1$, $\beta 2$.

2. Defina a taxa de aprendizado η.

3. Divida o conjunto de dados em três subconjuntos: Treinamento; Validação; e Teste;

4. Para *i* de 1 até *M* calcule

(a) Calcule os gradientes utilizando o conjunto de treinamento

(b) $\eta_0 = \eta * r$, onde r ~ U(0,1).

(c) Faça o ajuste dos pesos.

$$
\begin{aligned}
W1_{(i+1)} &= W1_{(i)} + \eta_0 \nabla W1_{(i)} \\
W2_{(i+1)} &= W2_{(i)} + \eta_0 \nabla W2_{(i)} \\
\beta 1_{(i+1)} &= \beta 1_{(i)} + \eta_0 \nabla \beta 1_{(i)} \\
\beta 2_{(i+1)} &= \beta 2_{(i)} + \eta_0 \nabla \beta 2_{(i)}
\end{aligned}
$$

(d) Calcule e armazene a função Desvio para os conjuntos de Treinamento e Validação

(e) Se desvio para o conjunto de treinamento for menor do que aquele calculado no passo anterior, mantenha a atualização dos pesos, caso contrário desfaça-a.

(f) Armazene os pesos com menor Desvio de Validação

5. Para o conjunto de Teste, calcule o Desvio com os pesos armazenados em 4(f)

Nesta seção, foram apresentados os componentes do algoritmo *backpropagation* para o ajuste das RNGs para a modelagem de dados, os quais a variável resposta pertença à família exponencial.

Na próxima seção, serão apresentados alguns resultados de comparações empíricas entre as RNGs e os modelos lineares generalizados, utilizando bases de dados reais.

4. Metodologia e Resultados

A avaliação do potencial das Redes Neurais Generalizadas é realizada através de comparações empíricas com modelos convencionais, neste caso, com os Modelos Lineares Generalizados - MLGs. Muitos problemas de predição de dados podem ser modelados através de MLGs. Deseja-se avaliar a capacidade das RNGs de retornar valores preditos melhores do que os resultados encontrados com os modelos convencionais. Sabe-se que a escolha da função de ligação influi no resultado do modelo e, portanto, a melhor alternativa consiste em comparar o modelo de RNGs com diversos MLGs, cada qual com uma função de ligação distinta. O modelo a ser escolhido será aquele que retornar o menor valor da *deviance* para os conjuntos de treinamento e teste. A função *deviance* é definida como a diferença entre a log-verossimilhança do modelo saturado e o modelo ajustado (NELDER e WEDDERBURN, 1972). De forma sucinta, maximizar a função log-verossimilhança é equivalente a minimizar a função *Deviance.*

No caso dos modelos RNGs, o valor da *deviance* é dada via Validação Cruzada. Ao utilizarmos a validação cruzada, a base de dados é subdividida aleatoriamente em três grupos: o primeiro, denominado grupo de Treinamento e contendo 60% do total de observações, é utilizado para o treinamento e/ou ajuste dos parâmetros da rede, o segundo, denominado grupo de Validação e contendo 20% dos registros, é utilizado como critério de parada do processo de treinamento e/ou escolha da solução final e o terceiro conjunto, denominado conjunto de Teste, é utilizado para a verificação do processo. Durante o treinamento da rede, a *deviance* é avaliada a cada atualização do vetor de pesos. Simultaneamente, o valor da *deviance* para o conjunto de validação também é calculado. O vetor de parâmetros final da rede corresponde àquele que representa o menor valor da *deviance* de validação durante a fase de treinamento. A Figura 4.1 ilustra o processo de treinamento e validação cruzada para a escolha do vetor final de parâmetros, aplicado ao bando de dados *Caranguejo.*

Para realizar uma comparação preliminar empírica, foram utilizadas duas bases de dados reais, descritas a seguir:

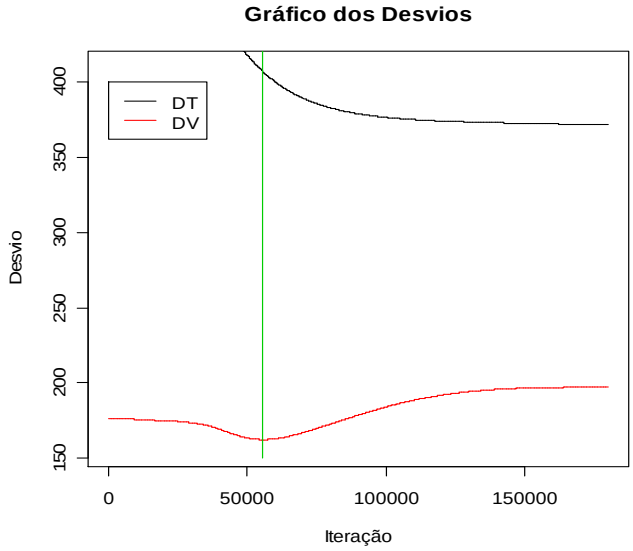
A primeira relaciona-se a um estudo (BROCKMANN, 1996), no qual deseja-se prever o número de *satélites* (ou seja, número de caranguejos machos que rodeiam uma fêmea) de acordo com as características dos 173 caranguejos fêmeas. As características são: cor, estado da espinha dorsal, peso e comprimento da carapaça, sendo que as duas primeiras variáveis são qualitativas e as outras duas quantitativas. A variável resposta, o número de satélites, neste caso segue a distribuição de *Poisson.*

A segunda base de dados representa uma pesquisa de *marketing* de uma determinada companhia telefônica dos Estados Unidos e apresenta informações de 1,000 domicílios e suas respectivas famílias. Porém, como muitas destas entrevistas apresentaram dados faltantes para uma ou mais variáveis, tais registros foram excluídos, restando um total de 757. Foram observadas 10 variáveis, dentre as quais apenas duas eram quantitativas (nº- de vezes em que o indivíduo se mudou nos últimos 10 anos e o uso médio mensal do domicílio). Os dados foram obtidos através de uma pesquisa realizada por telefone (WATSON, 1982). A variável de interesse neste estudo, que é a preferência por determinada operadora de telefonia, segue a distribuição Binomial.

O modelo de RNGs para a distribuição *Poisson* foi comparado com o modelo linear generalizado com função de ligação logarítmica. O modelo de RNGs para Binomial foi comparado com três outros modelos lineares generalizados: um com função de ligação logística, outro com função *probit* e o terceiro com função complemento *log-log*. Os resultados das comparações para a base de dados *Caranguejo (Poisson) e Companhia Telefônica (Binomial)* estão descritos a seguir. As RNGs utilizadas nos estudos empíricos realizados com dados reais tinham 10 neurônios na camada escondida, uma taxa de aprendizado de 0,00002 e realizavam no máximo 150,000 iterações para cada uma das 100 reamostragens.

POISSON

Figura 4.1 - Gráfico de comportamento da *deviance* para os conjuntos de treinamento e validação referente à base de dados Caranguejo.



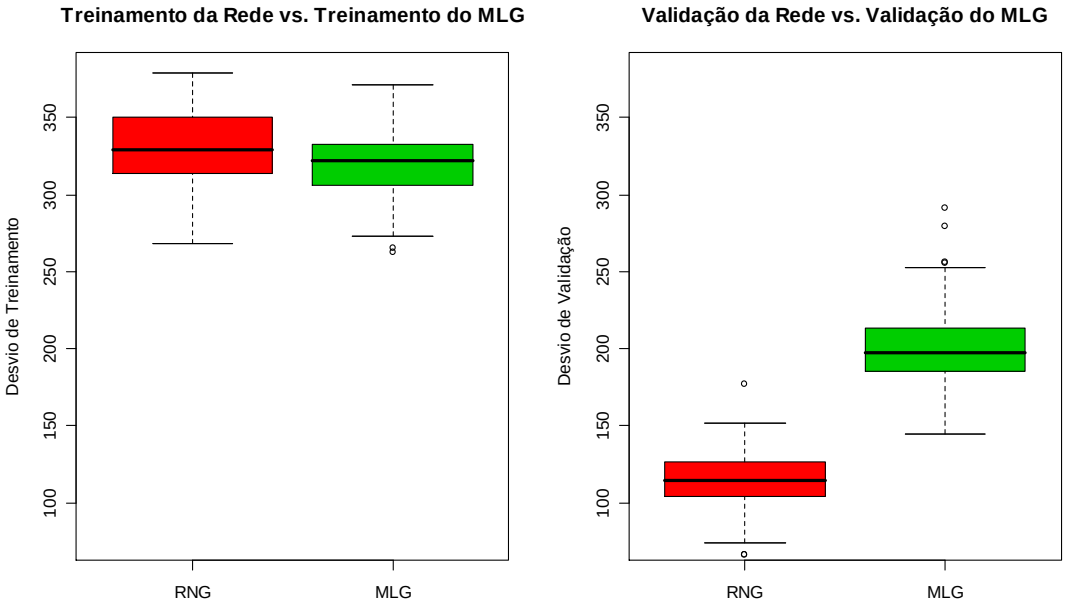
A Figura 4.1 mostra o número de iterações e as curvas de *deviance* para uma amostra dos conjuntos de treinamento e validação para a base de dados *Caranguejo*. A linha verde marca a iteração onde ocorreu a *deviance* mínima de validação, representando o ponto no qual foram definidos os parâmetros da rede.

De acordo com a Figura 4.1, não se faz necessário realizar todas as 150 mil iterações, pois com aproximadamente 55 mil o algoritmo encontrou uma *deviance* de validação mínima, ou seja, ele encontrou o vetor de parâmetros da rede que retorna o menor erro de predição.

Tabela 4.1 - Resultados da *deviance* para a base Caranguejo, utilizando a verossimilhança de *Poisson*

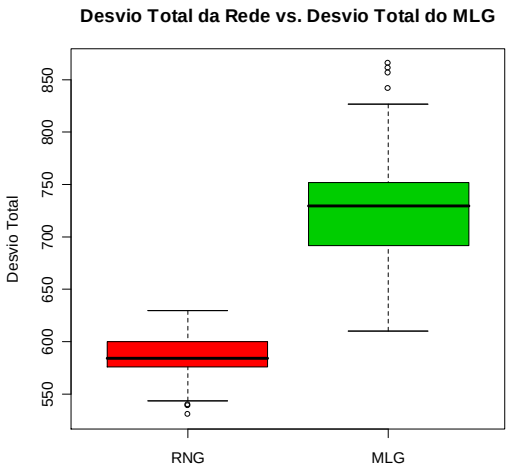
Modelo	Deviance	Média	Mediana	Desvio Padrão
RNG	Treinamento	331,2	329,5	24,48
	Validação	115,2	114,7	19,82
	Teste	139,2	138,5	22,36
	Total	585,6	584,1	19,07
MLG	Treinamento	319,5	322,1	22,95
	Validação	201,7	197,2	26,02
	Teste	202,5	196,3	26,45
	Total	723,7	729,3	50,44

Figura 4.2 - Gráfico comparativo do comportamento dos desvios de Treinamento e Validação entre a RNG e o MLG referente à base C*aranguejo*



Podemos observar pelos dados apresentados na Tabela 4.1 e visualmente na Figura 4.2 que mesmo o Modelo Linear Generalizado se ajustando melhor aos dados (amostra de treinamento) do que a Rede Neural Generalizada, esta possui resultados de previsão melhores do que o Modelo Linear Generalizado para os dados da base *Caranguejo*.

Figura 4.3 - Gráfico comparativo do comportamento do Desvio Total entre a RNG e o MLG referente à base C*aranguejo*



Na Figura 4.3, podemos observar que a *deviance* total (soma das *deviances* de treinamento, validação e de teste) da RNG é menor do que a do MLG, fato este que fornece

evidência da existência de interação não-aditiva ou de correlação não-linear entre as variáveis preditoras, entretanto essa característica não pode ser quantificada em função da sua representação interna na topologia da Rede Neural, o que faz dela um modelo preditivo.

BINOMIAL

Podemos observar pela Tabela 4.2 que a RNG modela melhor os dados provenientes da base *Companhia Telefônica,* a qual tem variável resposta regida pela distribuição binomial. Observamos também que os resultados da função desvio entre os vários modelos são muito próximos, o que evidencia a não existência de não-linearidade dos dados ou iteração entre as variáveis preditoras, evento contrário ao observado com relação à base de dados *Caranguejo* onde as *deviances* totais dos modelos são muito distintas, sendo obtida pela rede neural generalizada bem menor do que a obtida via modelo linear generalizado.

Tabela 4.2 Resultados da *deviance* para a base Companhia Telefônica, utilizando a verossimilhança da Binomial.

Modelo	Deviance	Média	Mediana	Desvio Padrão
RNG	Treinamento	581,1	580,2	11,93
	Validação	199,5	200,0	5,84
	Teste	201,7	200,8	8,97
	Total	982,4	979,1	14,50
MLG Logit	Treinamento	559,1	560,8	9,71
	Validação	216,1	211,7	18,59
	Teste	216,0	212,6	19,54
	Total	991,2	987,8	23,77
MLG Probit	Treinamento	559,5	561,1	9,86
	Validação	215,8	212,3	17,72
	Teste	216,0	212,3	19,36
	Total	991,3	988,6	22,43
MLG Complemento Log-log	Treinamento	559,7	560,9	10,25
	Validação	223,1	214,3	24,35
	Teste	224,6	214,0	29,47
	Total	1007,0	1009,0	32,84

5. Conclusão

Neste trabalho, o modelo de Redes Neurais Generalizadas foi apresentado. Este modelo utiliza a rede neural do tipo MLP agregada à estrutura de regressão dos Modelos Lineares Generalizados.

Os resultados obtidos ilustram a capacidade das RNGs de ajustar automaticamente seus parâmetros com base em um conjunto de observações sem uso de informações prévias sobre a correlação das variáveis de interesse, gerando modelos de predição com desempenho, no mínimo, comparáveis aos modelos tradicionais. Dessa forma, as RNGs são ferramentas semi-paramétricas que permitem validar os MLGs.

Esta característica evidencia o ganho obtido a partir da integração dos modelos estatísticos tradicionais com os modelos computacionais neurais (RNA). Esta integração vem sendo realizada em muitos trabalhos, principalmente na área da saúde (BIGANZOLI *et al.*, 1998 e 2002, e FORAGGI e SIMON, 1995). Porém, apesar de utilizarmos um algoritmo de treinamento (*backpropagation*) diferente do utilizado pelos autores anteriormente citados (NEWTON-RAPHSON), esta é a primeira vez em que são demonstradas as equações de ajuste dos pesos da rede, além de uma forma geral para o processamento de uma RNA com distribuição da família exponencial.

Vale salientar que os modelos de Redes Neurais Generalizadas são modelos preditivos e não explicativos. Ou seja, não é evidente o impacto da variação de uma determinada variável explicativa na variável resposta. Por outro lado, os modelos preditivos possuem melhor comportamento na estimação de dados, onde a variável resposta é faltante. É o caso, por exemplo, de séries temporais (ZHANG e QI, 2005).

Por ser um modelo semi-paramétrico estimado a partir de algoritmos iterativos, a estimação de intervalos de confiança para a variável resposta pode ser realizada a partir de métodos de reamostragem. Ao repetirmos o processo de ajuste dos parâmetros da rede, para cada reamostragem dos conjuntos de treinamento, validação e teste, obtém-se múltiplos valores preditos para cada elemento amostral. Pode-se, então, gerar intervalos de confiança para a saída do modelo a partir dessas distribuições empíricas.

Como trabalhos futuros, pretende-se: explorar as RNGs aplicando-se algoritmos de *pruning* para a simplificação do modelo neural (COSTA *et al.*,2003); adaptar algoritmos mais eficientes (mais rápidos) e mais robustos (menos sensíveis a mínimos locais) para a atualização dos pesos da rede e utilizar as RNGs como método preditivo para imputação de dados do IBGE.

Referencias Bibliográficas

Ambrogi, F.; Lama, N.; Boracchi, P. e Biganzoli, E. (2007). Selection of artificial neural network models for survival análisis with Genetic Algorithms, Computational Statistics & Data Analysis 52, 30-42.

Biganzoli, E. ; Boracchi, P.; Marubini, E. e Mariani,L. (1998) Feed forward Neural Networks for the analysis of censored survival data: A partial logistic regression approach. Statistics in Medicine 17 1169-1186

Biganzoli, E.; Boracchi, P. e Marubini, E. (2002) A general framework for neural network models on censored survival data .Neural Networks 15 209-218

Biganzoli, E. M.; Boracchi, P.; Ambrogi, F. e Marubini, E. (2006). Artificial neural network for the joint modeling of discrete cause-specific harzards, Artificial Intelligence in Medicine 37, 119-130.

Braga, A.P.; Carvalho, A.C.P.L.F e Ludermir, R.B. (2000) Redes Neurais Artificiais Teoria e Aplicações. LTC, Rio de Janeiro

Braga A.P.; Takahashi, R.; Costa M.A. e Teixeira, R. (2006). Multi-Objective Algorithms for Neural Networks Learning. Studies in Computational Intelligence. Springer.

Brockmann, H. J. (1996) Satellite male groups in horseshoe crabs. Limulus polyphemus. Ethology 102 1-21.

Costa, M.A.; Braga, A.P. e Menezes, B.R. (2003) Improving neural networks generalization with new constructive and pruning methods. Journal of Intelligent & Fuzzy Systems 13 75-83.

Costa, M. A. e Braga, A. P. (2006) Optimization of neural networks with multi-objective lasso algorithm. In IEEE World Congress on Computational Intelligence, Vancouver.

Cox, D. R. e Oakes, D. (1984). Analysis of Survival Data, London, Chapman andHall.

Cybenko, G. (1988) Continuos valued neural networks with two hidden layers are sufficient.Technical report. Department of Computer Science, Tufts University.

Cybenko, G. (1989) Approximation by superpositions of a sigmoid function. Mathematics of Control, Signals and Siystems, 2 303-314

Dugas, C; Bengio, Y; Chapados, N; Vicent, P; Denoncourt, G e Fournier, C (2003). Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking. In: Jain, L e Shapiro, A.F. (2003). Intelligent Techniques in the insurance industry: Theory and Applications. p.1-31

Fahlman, S.E. (1988) An empirical study of learning speed in back-propagation networks. Technical Report, Carnegie Mellow University

Faraggi, D. e Simon R. (1995) A neural network model for survival data. Statistics in Medicine 14 73-82
Haykin, S. (2001) Redes neurais: Princípios e prática. Bookman, Porto Alegre.

Lisboa, P. J. G.; Etchells, T. A.; Jarman, I. H.; Aung, M. S. H.; Chabaud, S.; Bachelot, T.;

Perol, D.; Gargi, T.; Bourdès, V.; Bonnevay, S. e Négrier, S. (2008). Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer, Neural Networks 21, 414-426.

Nelder, J. A. e Wedderburn, R.W.M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society A 135, 370-384.

Riedmiller, M. e Braun, H. (1993) A direct adaptive method for faster backpropagation learning: The Rprop algorithm. In. Proceedings of the IEEE Intl. Conf. on Neural Networks, pp 586-591, San Francisco, CA.

Rumelhart, D.E.; Hilton, G.E. e Willians, R.J. (1986) Learning Internal Representations of Back-propagation Errors. In. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, pp. 318-362, Cambridge, MA:MIT Press.

Silva, F.M. e Almeida, L.B. (1990) Speeding up backpropagation. In. Ekmiller,R., editor, Advanced Neural Computers, pp. 151-158, Amsterdam, Elsevier North Holland.

Tollenaere, T. (1990) SuperSAB: Fast adaptive back propagation with good scaling properties. Neural Networks, 3(5) 561-573

Watson, J.W.(1982) citado em Chambers, J.M. e Hastie, T.J. (1992) Statistical Models in S.S. Wadsworth and Brooks, Pacific Grove, CA , pág. 49.

Zhang, G. P. e Qi, M. (2005) Neural Network forecasting for seasonal and trend time series, European Journal of Operational Research 160, 501-514.

Zhang, W. e Zhang, X. (2008). Neural Network modeling of survival dynamics of homometabolous insects: A case study, Ecological Modelling 211, 433-443.

Abstract

This paper aims at presenting the generalized neural networks. These models add together the advantages of the likelihood structure of generalized linear models and the flexibility of artificial neural networks to model the nonlinear and nonadditive interactions between predictor and response variables. The training is done through the interactive descendent gradient method, which aims at minimizing the model deviance function, estimated by cross-validation. The results obtained so far show that the generalized neural networks are at least as good as the comparable generalized linear models.

Agradecimentos

Os autores agradecem à FAPEMIG, CNPq e ao IBGE pelo apoio financeiro.

O uso de modelos de séries temporais no estudo da produção de álcool no Brasil

Célia Mendes Carvalho Lopes[10]
Airlane Pereira Alenca[11]
Franco de Sá Barroso Lippi [12]
Flávio Hideki Yamamoto [13]

Resumo:

Nesse trabalho apresentamos resultados de análises de dados via técnicas de séries temporais. A evolução tecnológica industrial acarretou a necessidade de se organizar a produção através do estudo sobre previsão de demanda, de capacidade e de estoques de matéria-prima em um determinado período de tempo, caracterizando o estudo e análise das séries temporais na indústria. Neste trabalho será apresentada a forma como a sazonalidade e a tendência são identificadas em uma série temporal referente a dados de produção de álcool no Brasil, tendo como principal objetivo a previsão. Em particular, serão apresentados os resultados do ajuste e as previsões obtidas utilizando-se o método de suavização exponencial de *Holt-Winters* e modelos SARIMA.

Palavras - chave: Séries Temporais. Previsão. Método de *Holt-Winters*. SARIMA. Álcool.

10 Universidade Presbiteriana Mackenzie, Escola de Engenharia, Engenharia de Produção.
11 Universidade de São Paulo, Instituto de Matemática e Estatística, Departamento de Estatística.
12 Universidade Presbiteriana Mackenzie, Escola de Engenharia, Engenharia de Produção.
13 Universidade Presbiteriana Mackenzie, Escola de Engenharia, Engenharia de Produção.

1. Introdução

Antes da Revolução Industrial, que teve início no Século XVIII, a ação de se produzir algo era feita pelas mãos dos artesãos, os quais conduziam a produção do início ao fim, além de comercializar o produto final. Com o início da Revolução Industrial, a produção passou a ser realizada utilizando maquinário e passou a se produzir em larga escala, o que também ocorreu devido ao crescimento populacional, que demandou uma maior quantidade de produtos e mercadorias.

Esses aspectos acarretaram a necessidade de se organizar a produção por meio do estudo sobre a previsão de demanda, de capacidade e de estoques de matéria-prima em um determinado período de tempo, caracterizando o estudo e análise das séries temporais na indústria.

Uma série temporal é um conjunto de dados numéricos obtidos em uma sequência de instantes de tempo. A análise de séries temporais visa à formulação de modelos capazes de descrever a dependência dos dados com relação ao tempo e predizer valores futuros. Uma série temporal pode ser formada pelas componentes: tendência, sazonalidade, ciclo e termo aleatório (NASCIMENTO *et al*., 1996; LEVINE *et al.*, 2000).

Dados de séries temporais estão presentes em várias áreas do conhecimento, tais como: meteorologia (precipitação pluviométrica, altura de ondas no mar), economia (taxa mensal de desemprego, produção industrial) e epidemiologia (número mensal de novos casos de meningite).

Neste trabalho, será apresentada a forma como a sazonalidade e a tendência serão identificadas em uma série temporal referente a dados de produção mensal de álcool no Brasil de 1982 a 1989. Os dados serão analisados considerando o método de suavização exponencial de *Holt-Winters* e o ajuste do modelo SARIMA. O ajuste dos modelos será realizado considerando os dados até 1988 e serão calculadas previsões para a produção de álcool durante o ano de 1989. A fonte desses dados é a Agência Nacional de Petróleo - ANP e eles são disponibilizados na Internet pelo Instituto de Pesquisa Econômica Aplicada - IPEA (IPEADATA, 2007).

2. Séries Temporais

Uma série temporal é uma realização de um processo estocástico que consiste em uma sequência de variáveis aleatórias. Uma maneira tradicional de se analisar uma série temporal é por meio da sua decomposição nas componentes de tendência e sazonalidade. (MORETTIN e TOLOI, 2006).

Seja {Zt}, t = 0, 1, 2, ..., n, uma série temporal de interesse em que *n* é o número de observações coletadas. Segundo Morettin e Toloi (2006), os componentes tendência, sazonalidade e ciclo da série apresentada são dadas por:

Tendência (Tt em que t = 0, 1, 2,..., n): é caracterizada como um movimento regular e contínuo de longo prazo que pode ser crescente, decrescente ou constante. Em outras palavras, a tendência pode ser definida como a "direção" que a série temporal está apresentando.

Sazonalidade (St em que t = 0, 1, 2,..., n): corresponde às oscilações crescentes ou decrescentes que sempre ocorrem em um determinado período do dia, da semana, do mês ou do ano, como, por exemplo, quedas ou altas devido a safras de produtos agrícolas, a mudanças climáticas ou ao período de férias escolares.

Ciclo (Ct em que t = 0, 1, 2,..., n): corresponde às oscilações periódicas de longo prazo em torno da tendência, como, por exemplo, ciclos do comportamento de manchas solares que ocorrem a cada 11 anos.

Componente Aleatória (at em que t = 0, 1, 2,..., n): representa os movimentos aleatórios existentes nas séries de tempo e que não são previstos.

Assim, supondo um modelo aditivo, uma série temporal pode ser representada por:

$$Z_t = T_t + S_t + C_t + a_t .$$

Ao propor modelos para séries temporais, um conceito fundamental é a estacionariedade de um processo estocástico. Tal conceito é apresentado nas definições a seguir segundo Morettin e Toloi (2006).

Definição 1: Um processo estocástico $Z = \{ Z_t, t \in T \}$ diz-se estritamente estacionário se todas as distribuições finito-dimensionais permanecem as mesmas sob translações no tempo.

Na prática, é mais viável checar se a média e a variância de *Zt* permanecem constantes e se a covariância entre duas observações desse processo só depende da diferença entre os tempos, ou seja, temos um processo fracamente estacionário como definido a seguir, segundo Morettin e Toloi (2006).

Definição 2: Um processo estocástico $Z = \{ Z_t, t \in T \}$ diz-se fracamente estacionário ou estacionário de segunda ordem se, e somente se:

$E\{ Z_t \} = \mu_t = \mu$, constante, para todo $t \in T$;

$E\{ Z_t^2 \} < \infty$, para todo $t \in T$;

$\gamma(t_1, t_2) = cov\{ Z_{t_1}, Z_{t_2} \}$ é uma função de $|t_1 - t_2|$.

Segundo Russo *et al*. (2006 *apud* BOX *et al.*,1994), a autocorrelação é uma medida de dependência entre observações da mesma série separadas por um determinado intervalo de tempo chamado retardo ou defasagem. A autocorrelação ajuda no entendimento do comportamento da série, devido à necessidade de conhecer a relação existente entre os dados presentes e passados. Ao propor modelos, como o modelo SARIMA apresentado na seção 2.2, a autocorrelação é importante na identificação do modelo a ser ajustado. Além disso, tais modelos incluem um componente aleatório (a_t), denominado erro do modelo, e, em geral, supõe-se que esse componente é um processo estocástico de média zero, variância constante e não correlacionado. Tais processos são denominados ruído branco. Após o ajuste de tais modelos é necessário checar se a série dos resíduos do modelo (estimativas do erro) se comporta como um ruído branco, e nesse caso, deve ser verificado se as autocorrelações dos resíduos são estatisticamente não significantes.

2.1 Método de Suavização Exponencial Sazonal de *Holt-Winters*

Segundo Morettin e Toloi (2006), no método de *Holt-Winters* existem dois tipos de procedimento para modelar a sazonalidade, o aditivo e o multiplicativo. Esses procedimentos são baseados em três equações com constantes de suavização diferentes, que são associadas a cada uma das componentes do padrão da série: nível (μ_t), tendência (T_t) e sazonalidade (S_t).

Para séries que apresentam um efeito sazonal maior, quanto maior o nível da série, o que ocorre bastante na prática, podemos utilizar o método de suavização de *Holt-Winters,* proposto em Holt (1957) e Winters (1960), considerando um modelo com componente sazonal multiplicativa com período s:

$$Z_t = (\mu_0 + tT_0)S_t + a_t \text{, t = 1,...,N,}$$

de modo que as componentes sazonais são tais que $\sum_{t=1}^{s} S_t = s$.

As equações de suavização para as três componentes para $t = s+1,...,N$ são

$$\hat{\mu}_t = \alpha\left(\frac{Z_t}{\hat{S}_{t-s}}\right) + (1-\alpha)(\hat{\mu}_{t-1} + \hat{T}_{t-1}),$$

$$\hat{T}_t = \beta(\hat{\mu}_t - \hat{\mu}_{t-1}) + (1-\beta)\hat{T}_{t-1},$$

$$\hat{S}_t = \gamma\left(\frac{Z_t}{\hat{\mu}_t}\right) + (1-\gamma)\hat{S}_{t-s},$$

em que os parâmetros α, β e γ pertencem ao intervalo (0,1) e podem ser estimados de modo a minimizar a soma dos quadrados dos erros de previsão (resíduos ao quadrado), $\sum_{t=s+1}^{T}\left[Z_t - (\hat{\mu}_t + \hat{T}_t)\hat{S}_t\right]^2$.

O método de suavização exponencial de *Holt-Winters* com componente sazonal aditiva e a interpretação das equações e obtenção dos valores iniciais de cada componente são apresentadas com exemplos em *Winters* (1960) e detalhadamente em Montgomery e Johnson (1976).

A previsão para a variável no instante *t* + *h*, considerando as observações até o instante *t*, são:

$\hat{Z}_t(h) = (\hat{\mu}_t + h\hat{T}_t)\hat{S}_{t+h-s}$, *h* = 1, 2, ..., s,

$\hat{Z}_t(h) = (\hat{\mu}_t + h\hat{T}_t)\hat{S}_{t+h-2s}$, *h* = s+1, ..., 2s,

e assim sucessivamente.

A variância dos erros de previsão e a construção de intervalos de confiança para a previsão podem ser encontradas, por exemplo, em Montgomery e Johnson (1976) e em Sweet (1985).

2.2 SARIMA

De acordo com Werner e Ribeiro (2003), os modelos estacionários são aqueles que assumem que o processo está em equilíbrio. A estacionariedade é bastante importante na análise de uma série temporal, pois se sua média, variância e, de modo mais geral, sua

distribuição muda ao longo do tempo, é impraticável ajustarmos um modelo já que dispomos de somente uma série temporal.

Uma classe de modelos paramétricos muito utilizada na prática para a análise de séries temporais é a classe de modelos apresentada em Box e Jenkins (1970). Esses modelos assumem que ao calcularmos diferenças da série (Yt = □Zt = (1 –L) Zt = Zt – Zt-1, com L denotando o operador defasagem) obtemos processos estacionários, como apresentado na Definição 2. Uma série que vem de um processo estacionário deve apresentar valores em torno de uma média constante e tais valores devem permanecer ao redor de uma linha horizontal ao longo do tempo, indicando que a variância é constante ao longo do tempo. Para verificar se uma série é estacionária, seu comportamento ao longo do tempo deve ser analisado graficamente. Então, a primeira etapa da modelagem consiste na obtenção de uma série estacionária, o que pode incluir, além do cálculo de diferenças, outras transformações da série original, como, por exemplo, a transformação logarítmica.

Após obter uma série estacionária, é estudada a função de autocorrelação amostral (correlograma) da série estacionária com o objetivo de ajustar um modelo: autorregressivo (AR) (sob algumas condições de estacionariedade) em que a série de dados históricos é descrita por seus valores passados e pelo ruído branco a_t; médias móveis (MA), para o qual a série estacionária resulta da combinação de ruídos brancos a_t do período atual e de períodos anteriores; o autorregressivo e médias móveis (ARMA) que incluem componentes de um modelo AR e componentes de um modelo MA. De modo geral, temos o modelo ARMA(p, q), que é composto por *p* termos autorregressivos *(Yt-k, k = 1,...,p)* e *q* termos médias móveis (at-*k, k = 1,...,q)*, o que pode exigir um menor número de parâmetros do que ao considerarmos somente os modelos puramente AR ou puramente MA.

No caso de a média da série original não permanecer constante e ser necessário o cálculo de diferenças sucessivas até que a série se torne estacionária, denominamos ordem de integração (d) o número de diferenças necessárias para tornar a série estacionária. Ao ajustarmos um modelo ARMA(p,q) para uma série que precisou de *d* diferenças, denominamos o modelo ajustado de modelo ARIMA(p,d,q).

Segundo Werner e Ribeiro (2003), os modelos ARIMA exploram a autocorrelação entre os valores da série em instantes sucessivos, mas quando os dados são observados em períodos superiores a um ano, a série também pode apresentar autocorrelação para uma estação de sazonalidade de período *s.* Um dos modelos que contemplam as séries que apresentam componente sazonal são conhecidos como SARIMA. Os modelos SARIMA contêm uma parte não sazonal (ARIMA), com parâmetros (p,d,q), e uma sazonal, com parâmetros (P,D,Q)*s*. O modelo mais geral é dado pelas equações:

$(1-\phi_1 L-...-\phi_p L^p)(1-\Phi_1 L^s-...-\Phi_P L^{sP})\ (1-L)^d (1-L^s)^D Z_t = \phi_0 + (1-\theta_1 L-...-\theta_q L^q)$

$(1-\Theta_1 L^s-...-\Theta_Q L^{sQ})a_t$ em que:

$(1-\phi_1 L-...-\phi_p L^p)$ é a parte autorregressiva não-sazonal de ordem *p;*

$(1-\Phi_1 L^s-...-\Phi_P L^{sP})$ é a parte autorregressiva sazonal de ordem *P* e estação sazonal *s*;

$(1-L)^d$ é parte de integração não-sazonal de ordem *d*;

$(1-L^s)^D$ é parte de integração sazonal de ordem *D* e estação sazonal *s*;

ϕ_0 é o intercepto do modelo;

$(1-\theta_1 L-...-\theta_q L^q)$ é a parte não-sazonal de médias móveis de ordem q;

$(1-\Theta_1 L^s-...-\Theta_Q L^{sQ})$ é a parte sazonal de médias móveis de ordem *Q* e estação sazonal *s*;

e a_t tem média zero, variância constante e são não correlacionados, ou seja, a série de erros a_t é um ruído branco.

Para séries com componente sazonal, pode ser necessário considerar diferenças sazonais. Por exemplo, para dados mensais com *s* = 12, pode ser necessário calcular 12 diferenças sazonais, $\left(1-L^{12}\right)Z_t = Z_t - Z_{t-12}$, para obter uma série estacionária. Após o cálculo de diferenças nas séries com o objetivo de obter uma série com uma média estável, a escolha do modelo SARIMA baseia-se na análise do correlograma. Outra ferramenta que auxilia a identificação de modelos ARIMA e SARIMA é o gráfico das autocorrelações parciais. Como, para a produção de álcool, esse gráfico não permitiu a clara identificação dos valores de *p, q*, *P* e *Q*, não o apresentamos nesse trabalho. Informações sobre a função de autocorrelação parcial e seu estimador são apresentados, por exemplo, em Morettin e Toloi (2006).

Assumindo que os erros a_t apresentam uma determinada distribuição de probabilidades, os parâmetros do modelo podem ser estimados pelo método de máxima verossimilhança e os estimadores apresentam assintoticamente distribuição normal. O cálculo da verossimilhança exata pode ser realizado a partir da representação do modelo SARIMA utilizando-se o modelo espaço de estados e, desse modo, a componente aleatória (at) pode ser estimada utilizando-se o Filtro de *Kalman* (DURBIN e KOOPMAN, 2001). No caso da série de produção de álcool, foi considerado que os erros têm distribuição normal. É importante checar a validade das suposições do modelo, verificando se os resíduos são ruídos branco

gaussiano, para que sejam válidos os testes de hipóteses e intervalos de confiança para a previsão como apresentado em Morettin e Toloi (2006). Para verificar se os resíduos são não correlacionados, o teste *Ljung-Box* pode ser utilizado para testar se todas as autocorrelações até uma determinada defasagem de interesse são nulas (LJUNG e BOX, 1978).

3. Série de produção de álcool

O conjunto de dados que utilizamos consiste na série de produção mensal de álcool, medidos em quantidade média de barris fabricados por dia, de janeiro de 1982 até dezembro de 1989. Os dados foram obtidos no *site* governamental IPEADATA (2007) e estão apresentados na Figura 1.

A partir dos dados até dezembro de 1988, será realizado um estudo sobre a previsão da produção de álcool em 1989, utilizando o método de suavização de *Holt-Winters* e o ajuste de um modelo SARIMA para posterior comparação entre as previsões e os dados que foram observados em 1989. Para avaliar quais métodos de previsão estão adequados para essa série, devemos também analisar os resíduos obtidos segundo os dois métodos.

Figura 1: Série de produção de álcool

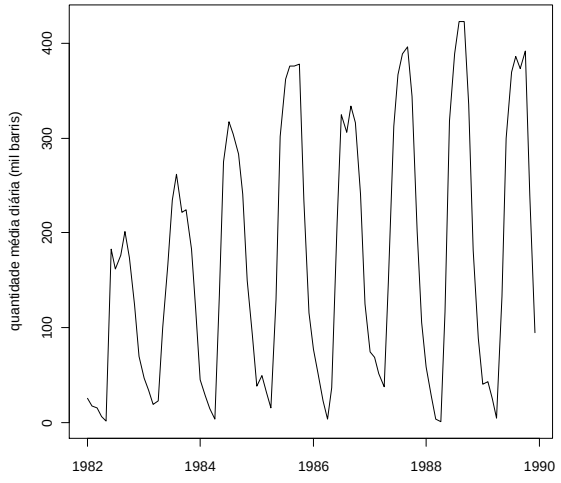
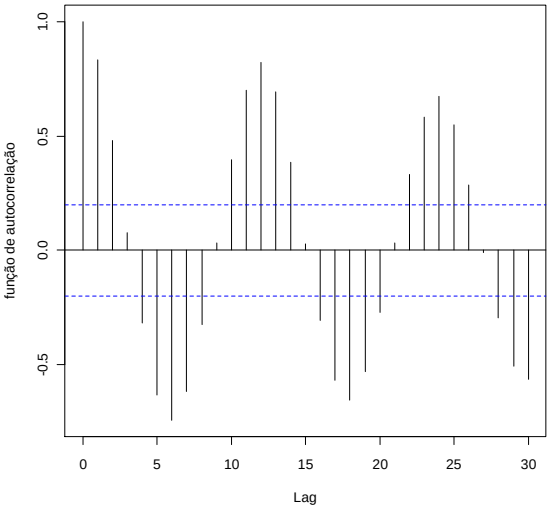


Figura 2: Função de autocorrelação amostral da série de produção de álcool



Na Figura 1, observa-se, claramente, que a produção de álcool apresenta uma sazonalidade de 12 meses. Tal sazonalidade é esperada devido às safras da cana-de-açúcar. Em média, a produção de açúcar aumenta ao longo do tempo e a componente sazonal também parece aumentar, apesar de oscilar bastante.

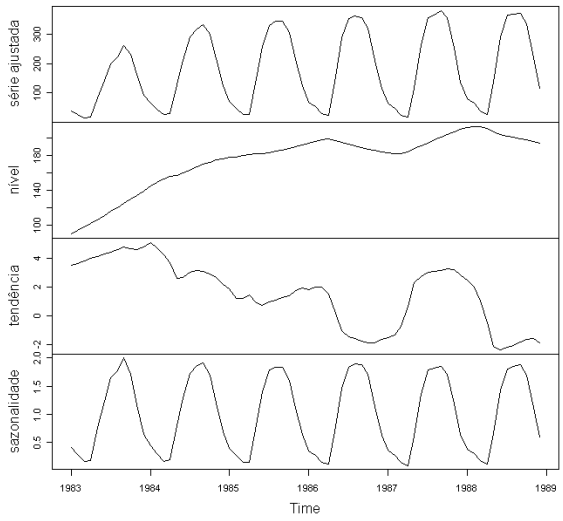
Na Figura 2, o correlograma da série da produção de álcool apresenta maiores autocorrelações para as defasagens (Lags) iniciais, por exemplo, 1 e 2. Além disso, o correlograma apresenta um comportamento cíclico com um maior valor de autocorrelação para a defasagem 12, ou seja, temos maiores autocorrelações entre as observações *Zt* e *Zt*-12. Esse comportamento é esperado para séries com sazonalidade anual.

3.1 Análise via *Holt-Winters*

A série da produção de álcool foi analisada segundo o método de *Holt-Winters*, utilizando-se a função *Holt-Winters* (MEYER, 2002) da biblioteca do programa gratuito *R*.

No método de Holt-Winters, a série pode ser decomposta nas componentes nível, tendência e sazonalidade, como explicado na seção 2.1. As estimativas dos parâmetros α, β e γ foram, respectivamente, 0,0171; 0,4706 e 0,2885. Na Figura 3, são apresentadas as séries correspondentes às três componentes do modelo no período de 1983 a 1988, já que os dados do primeiro ano são considerados para inicializar a suavização. A componente do nível apresenta um aumento ao longo do tempo, a tendência apresenta valores mais baixos entre junho de 1986 e março de 1987 e a componente sazonal apresenta os maiores valores nos meses de março a novembro.

Figura 3: Série ajustada e suas componentes obtidas pelo método de *Holt-Winters*.



A Figura 4 apresenta os valores observados da produção de álcool de 1982 a 1989, a série ajustada para o período de 1983 a 1988 e os valores previstos para 1989. As séries observada e ajustada apresentam valores semelhantes, mas as maiores discrepâncias ocorrem nos meses de maior produção de álcool, de junho a outubro. Já as previsões subestimam consideravelmente a produção de álcool de maio a novembro de 1989.

A Figura 5 apresenta o gráfico dos resíduos obtidos após o ajuste via método de *Holt-Winters* ao longo do tempo, o correlograma dos resíduos, gráfico de dispersão dos resíduos em função dos valores ajustados e o gráfico quantil-quantil dos resíduos. A série dos resíduos não se comporta como um ruído branco, pois os resíduos não estão dispersos de modo aleatório ao longo do tempo e há autocorrelações significativas no correlograma. No gráfico dos resíduos em função do valor ajustado, podemos observar que a variância dos resíduos é maior, quanto maior o valor ajustado, indicando que, provavelmente nesses casos, há maior variabilidade do erro de previsão. Observando o gráfico quantil-quantil, podemos observar que os resíduos parecem apresentar distribuição normal. É importante frisar que o método *Holt-Winters* não requer que os erros sejam gaussianos e se comportem como um ruído branco, mas a avaliação dos resíduos pode auxiliar a comparação entre os métodos de previsão.

A soma dos resíduos ao quadrado foi igual a 85359,4, considerando o período de 1983 a 1988. A soma dos erros de previsão para 1989 foi igual a 23028,0.

Figura 4: Séries observada, ajustada e prevista pelo método de *Holt-Winters*

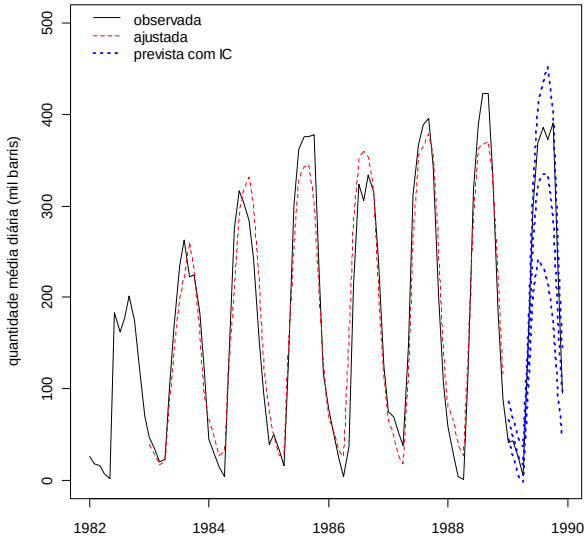
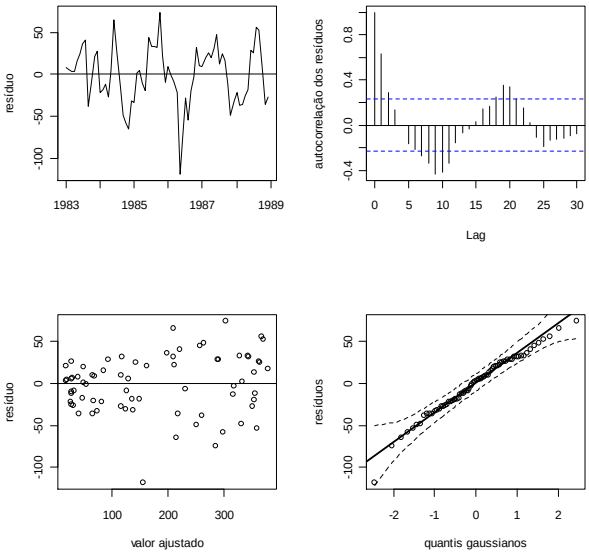


Figura 5: Resíduos obtidos pelo método de *Holt-Winters*



3.2 Análise via ajuste do Modelo SARIMA

Após o cálculo da diferença de ordem 12 da série de produção de álcool, a série parece estacionária e podemos propor um modelo estacionário ARMA incluindo a parte sazonal. Foi então ajustado um modelo SARIMA(1,0,1)x(1,1,1)12 , utilizando o comando Arima da biblioteca *forecast* do programa *R*. É importante observar que utilizando esse comando, as estimativas dos coeficientes dos termos médias móveis estão multiplicadas por

(-1), considerando a parametrização usual apresentada na seção 2.2. Além disso, a estimação é realizada por máxima verossimilhança exata, utilizando a representação do modelo ARIMA via modelo espaço de estados (DURBIN e KOOPMAN, 2001). Foi realizada uma análise dos resíduos obtidos a partir do ajuste do modelo SARIMA(1,0,1)x(1,1,1)12 , e todas as suposições do modelo parecem satisfeitas (apresentaremos a análise de resíduos somente para o modelo final, já que os resultados são muito semelhantes).

De acordo com os resultados apresentados na Tabela 1, esse modelo pode ser simplificado, já que apresenta parâmetros não significativos, considerando-se o nível de significância de 5%. Retirando-se o termo correspondente ao AR sazonal e MA não sazonal, o modelo final obtido é um SARIMA(1,0,0)x(0,1,1)12, dado por

$$(1-\phi_1 L)\,(1-L^{12})Z_t = \phi_0 + (1-\Theta_1 L^{12})a_t ,$$

em que ϕ_0 é o intercepto do modelo, ϕ_1 é o coeficiente do termo AR(1), Θ_1 é o coeficiente do termo MA(1) sazonal (SMA1).

As estimativas, erros padrões, estatísticas-*z* e *p*-valores correspondentes a esse modelo estão na Tabela 2.

Tabela 1: Modelo SARIMA(1,0,1)x(1,1,1)12 proposto para série de produção de álcool

Coeficiente	Estimativa	Erro Padrão	Estat.z	p-valor
ar1	0,6467	0,0168	4,9857	<0,0001
ma1	-0,1083	0,0319	0,6066	0,5441
sar12	0,0495	0,0664	0,1920	0,8478
sma12	0,5448	0,0614	-2,1979	0,0280
intercepto	1,2831	0,3225	2,2593	0,0239

Tabela 2: Modelo final SARIMA(1,0,1)x(1,1,1)12 proposto para série de produção de álcool

Coeficiente	Estimativa	Erro Padrão	Estat.z	p-valor
ar1	0,6984	0,0072	8,2242	<0,0001
sma12	0,4764	0,0185	-3,5031	0,0005
intercepto	1,2711	0,3898	2,0360	0,0417

Os valores ajustados para a produção de álcool até 1988 e previsões para 1989 e seus intervalos de confiança estão apresentados na Figura 6. Os valores observados em 1989 são próximos dos previstos, indicando que o modelo consegue prever bem a produção.

A Figura 7 apresenta os gráficos de resíduos para os dados de produção de álcool,

ajustados via modelo SARIMA. Os resíduos parecem apresentar média e variância constantes, já que no gráfico dos resíduos ao longo do tempo e em função dos valores ajustados, observa-se que os resíduos estão sempre em torno do zero e com variância aproximadamente constante. Não há evidências para rejeitar que as autocorrelações até a defasagem 12 sejam todas nulas, pois o nível descritivo do teste de *Ljung-Box* foi igual a 0,3747. O comportamento da autocorrelação dos resíduos também pode ser observado no correlograma. A partir destes resultados, podemos concluir que os resíduos se comportam como um ruído branco. Além disso, os resíduos apresentam distribuição próxima da distribuição normal, como pode ser observado no gráfico quantil-quantil.

Figura 6: Séries observada, ajustada e prevista pelo ajuste do modelo final SARIMA(1,0,0)x(0,1,1)$_{12}$

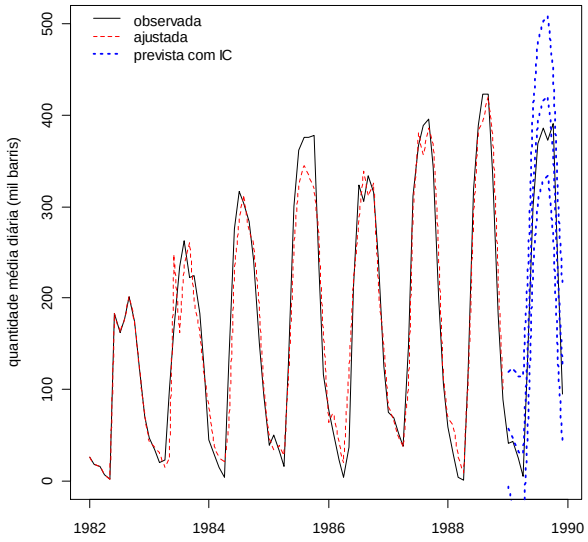
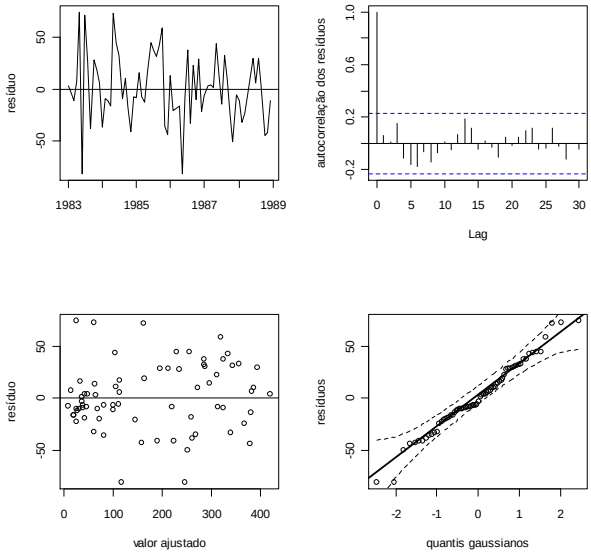


Figura 7: Resíduos obtidos pelo modelo SARIMA(1,0,0)x(0,1,1)$_{12}$



A soma dos resíduos ao quadrado foi igual a 73337,1, considerando o período de 1983 a 1988. A soma dos erros de previsão para 1989 foi igual a 7281,9.

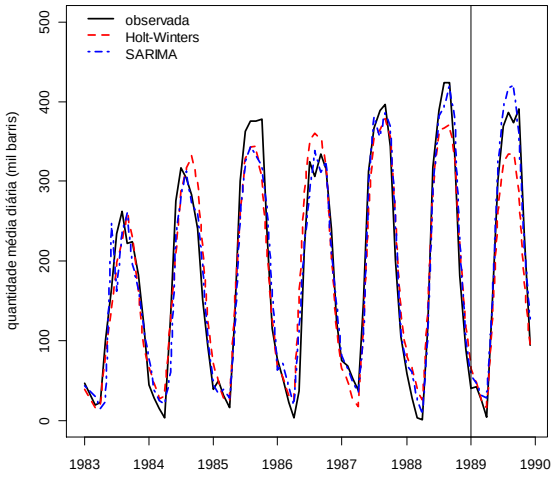
4. Comparação entre os resultados obtidos, segundo as duas metodologias e conclusões

Para escolher a melhor metodologia para prever a produção de álcool, podem ser analisados: o ajuste do modelo no período 1983-1988, os erros de previsão e os tamanhos dos intervalos de confiança para as previsões para 1989.

Com respeito ao ajuste do modelo, a soma dos resíduos ao quadrado no período 1983-1988 é menor para o modelo SARIMA (73337,1) do que para o método de *Holt-Winters*

(85359,4). Além disso, os resíduos do modelo SARIMA se comportam como um ruído branco gaussiano, de acordo com as suposições do modelo. Na Figura 8, é possível observar que os valores ajustados obtidos pelo modelo SARIMA são mais próximos dos valores observados para os meses com maior produção de álcool do que os valores ajustados pelo método de *Holt-Winters.*

Figura 8: Valores observados para a produção de álcool de 1983-1989 e valores ajustados de 1983-1988 e previsões para 1989, segundo os métodos *Holt-Winters* e SARIMA.



Também na Figura 8, podemos observar que as previsões obtidas a partir do modelo SARIMA superestimam a produção de álcool, enquanto as previsões obtidas pelo método de *Holt-Winters* subestimam a produção de álcool. Também nota-se que as previsões obtidas pelo modelo SARIMA estão mais próximas dos valores observados do que as obtidas pelo *Holt-Winters*. Isso resultou em uma menor soma dos erros de previsão ao quadrado para o modelo SARIMA (7281,9) do que para o método de *Holt-Winters* (23028,0).

Na Tabela 3, são apresentadas as previsões para a produção de álcool mensal para 1989, com os respectivos intervalos de confiança com coeficiente de confiança de 95% segundo ambos os métodos. Observa-se que, utilizando-se o modelo SARIMA, os intervalos de confiança para a previsão são bem maiores do que os obtidos pelo método de *Holt-Winters*, a amplitude dos intervalos aumenta com o tempo (como esperado para o modelo SARIMA) e o limite inferior do intervalo é até negativo.

Tab.3:Valor observado da produção de álcool em 1989 e previsão, intervalo de confiança para a previsão com coeficiente de confiança 95% e amplitude de cada intervalo para 1989 segundo cada método

Mês	Valor observado	Holt-Winters				SARIMA			
		Previsão	Lim.Inf.	Lim.Sup	Amplitude	Previsão	Lim.Inf.	Lim.Sup	Amplitude
jan-89	41	65,8	45,4	86,2	40,8	46,5	-17,4	110,4	127,8
fev-89	43	47,6	27,2	68,1	40,9	29,5	-51,6	110,6	162,2
mar-89	25	25,4	4,9	45,8	40,9	9,8	-79,8	99,4	179,1
abr-89	5	16,8	-3,7	37,3	41,0	3,9	-90,3	98,0	188,3
mai-89	133	114,4	87,0	141,7	54,8	109,5	12,9	206,2	193,3
jun-89	300	261,9	204,7	319,1	114,4	295,0	196,9	393,1	196,2
jul-89	369	324,8	241,1	408,5	167,4	365,8	266,8	464,7	197,9
ago-89	386	334,7	233,2	436,1	202,9	389,4	290,0	488,8	198,8
set-89	373	334,7	217,3	452,1	234,8	393,7	294,0	493,4	199,4
out-89	391	287,6	172,1	403,0	231,0	330,1	230,3	430,0	199,7
nov-89	239	178,2	96,1	260,4	164,3	193,2	93,2	293,1	199,9
dez-89	95	92,8	46,2	139,5	93,3	97,8	-2,2	197,8	200,0

Outra metodologia para a estimação de modelos semelhantes aos considerados nos métodos de suavização simples, de Holt e Holt-Winters, considera modelos espaço de estados não lineares como apresentado em Ord *et al*. (1997) e em Hyndman *et al*. (2002). A metodologia proposta em Hyndman *et al*. (2002) está implementada no pacote *R* e os resultados podem ser obtidos usando o comando *hw* do pacote *forecast*. Para os dados da produção de álcool, considerando um modelo espaço de estados não linear semelhante ao da metodologia *Holt-Winters* com sazonalidade multiplicativa (mas há diferenças comentadas em Hyndman *et al*. (2002)), a soma de quadrados dos erros de previsão para 1989 foi maior (33503,6) do que os apresentados nesse trabalho. Alternativamente, também poderíamos considerar o método de suavização de *Holt-Winters* com componente sazonal aditivo, apesar de o gráfico da série de produção de álcool indicar que a componente da sazonalidade deve ser multiplicativa. Considerando a sazonalidade aditiva, o ajuste é bem pior e a soma de quadrados dos erros de previsão (103072,2) é bem maior do que os obtidos pelos métodos apresentados nesse trabalho.

Com base nos resultados apresentados para ambas as metodologias consideradas neste trabalho, podemos concluir que o ajuste do modelo SARIMA é melhor para a obtenção de previsões da produção de álcool, já que apresentou melhor ajuste no período 1983-1989 (menor soma de resíduos ao quadrado), a suposição de que a série dos erros é um ruído branco com distribuição normal parece estar satisfeita e esse modelo apresentou menores erros de previsão em 1989. É importante comentar que esses resultados não podem ser generalizados, afirmando-se que um método sempre fornece melhores previsões do que o outro, o que requer uma análise detalhada para cada caso.

5. Considerações finais

A realização de previsões de demanda consiste em uma atividade extremamente importante em uma organização, pois pode revelar as tendências de mercado e, assim, contribuir no planejamento estratégico da empresa, além de auxiliar na solução de problemas imediatos que ocorrem em setores primordiais de uma empresa industrial, tais como: produção, estocagem, logística e vendas.

A análise de séries temporais vem sendo bastante utilizada no auxílio à tomada de decisão e planejamento da produção, buscando a minimização desse erro de previsão em relação à demanda, para assim favorecer as estratégias da empresa e diminuir custos de produção. Isso possibilita a obtenção de resultados economicamente favoráveis, além de se caracterizar como uma proteção, já que permite a realização de ações preventivas contra as mudanças bruscas do mercado em que a organização está inserida.

A aplicação das metodologias utilizadas requer atenção e prática do pesquisador para a escolha do modelo que melhor se ajuste ao comportamento dos dados ao longo do tempo.

No estudo apresentado neste artigo, o método de suavização exponencial de *Holt-Winters* e o ajuste do modelo SARIMA se mostraram adequados para a análise da produção de álcool no Brasil. Os dois modelos apresentaram resultados satisfatórios, mas o modelo SARIMA apresentou melhor ajuste e menores erros de previsão.

Finalmente, podemos concluir que a análise de séries temporais pode ser amplamente aplicada em dados industriais, mostrando ser uma ferramenta eficaz no auxílio à tomada de decisão e planejamento estratégico e gestão de produção, além de fornecer um melhor controle da utilização dos recursos disponíveis para a cadeia produtiva das empresas que buscam uma posição competitiva no mercado.

Referências bibliográficas

BOX, G. E. P.; JENKINS, G.M. Times series analysis: forecasting and control. 3.ed. São Francisco: Holden-Day, 1970. 553p. (Edição revisada, 1976).

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. Times series analysis: forecasting and control. 3.ed. São Francisco: Holden-Day, 1994. 592p.

DURBIN, J.; KOOPMAN, S. J. Time Series Analysis by State Space Methods. Oxford: Oxford University Press, 2001. 253p.

HYNDMAN, R. J.; KOEHLER, A. B.; SNYDER, R. D.; GROSE, S. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting, v. 18, n.3, p.438-454, 2002.

HOLT, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum, Carnigie Institute, v. 52, 1957.

IPEADATA - INSTITUTO DE PESQUISA ECONÔMICA APLICADA. Ministério do Planejamento, Orçamento e Gestão. Séries Históricas. Disponível em: <http://www.ipeadata.gov.br/ipeaweb.dll/ipeadata?77645578>. Acesso em 13 mar. 2007.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. Estatística: teoria e aplicações usando microsoft® excel em português. Rio de Janeiro: Livros Técnicos e Científicos Editora, 2000. 811p.

LJUNG, G. M.; BOX, G. E. P. On a measure of lack of fit in time series models. *Biometrika, v.* 65, p.553–564, 1978.

MEYER, D. Naive Time Series Forecasting Methods. R-News, v.2, n.2, p.7-10, 2002.

MONTGOMERY, D.C.; JOHNSON, L. A. Forecasting and time series analysis. New York: McGraw-Hill, 1976. 304p.

MORETTIN, **P. A.; TOLOI, C. M. C. Análise de séries temporais. 2. ed. São Paulo: Edgard Blücher, 2006. 535p. (Coleção Projeto Fisher).**

NASCIMENTO, N. O.; NAGHETTINI M.; HELLER L.; VON SPERLING, M. Investigação científica em engenharia sanitária e ambiental. Parte III: A análise estatística de dados e de modelos. Revista Engenharia Sanitária e Ambiental, Ano I, v. 4, p. 152-168, 1996.

ORD, J. K.; KOEHLER, A. B.; SNYDER, R. D. Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models. Journal of the American Statistical Association, v.. 92, n. 440, p. 1621- 1629, 1997.

SWEET, A. L. Computing the Variance of the Forecast Error for the Holt-Winters Seasonal Models. Journal of Forecasting, v. 4, n. 2, p. 235-243, 1985.

WERNER, L.; RIBEIRO, J. L. D. Demand forecasting: an application of the Box-Jenkins models in the technical assistance of personal computer. Gestão & Produção, v. 10, n. 1, p. 47-67, 2003.

WINTERS, P. R. Forecasting sales by exponentially weighted moving averages. Management Science, n. 6, p. 324–342, 1960.

[m1]

Abstract

In this article we present some results of data analysis via time series techniques. Industrial Technological developments led us to the need of organizing the production through a study on estimates of demand, capacity and stocks of raw material in a given period of time, featuring the analysis of time series in the industry. In this work will be shown how the seasonality and trend are identified in a time series data concerning the production of alcohol in Brazil, where the main goal is forecasting. In particular, the fitness results and forecasts are presented using the Holt-Winters exponential smoothing method and SARIMA models.

Keywords: Time Series. Forecast. Holt-Winters method. SARIMA. Alcohol.

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística Aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação, citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará, também, artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográfico e econométrico.

Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores seja brasileiro e nem resida no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo ‘duplo cego’, isto é, os artigos são avaliados sem a identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

O processo editorial da RBEs é eletrônico. Os artigos devem ser submetidos via *e-mail* para: rbe@ibge.gov.br.

Após a submissão, o autor receberá um código para acompanhar o processo de avaliação do artigo. Caso não receba um aviso com este código no prazo de uma semana, fazer contato com a secretaria da revista no endereço:

Revista Brasileira de Estatística – RBEs
ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - IBGE
Rua André Cavalcanti, 106, sala 111
Centro, Rio de Janeiro – RJ
CEP: 20031-170
Tels.: 55 21 2142-4682 (Sandra Cavalcanti Barros – Secretária)
55 21 2142-4686 (Ismenia Blavatsky – Editor–Executivo)
Fax: 55 21 2142-0501

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais enviados para publicação devem obedecer às normas seguintes:

1. Podem ser submetidos originais processados pelo editor de texto *Word for Windows* ou originais processados em LaTeX (ou equivalente) desde que estes últimos sejam encaminhados e acompanhados de versões em pdf, conforme descrito no item 3, a seguir;

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, se for o caso de constarem no documento, também devem figurar nesta página;

3. No caso de a submissão não ser em *Word for Windows*, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, LaTeX). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página, como descrito no item 2, e outro contendo apenas o título, sem a identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;

4. A segunda página do original deve conter resumos em português e inglês (*abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser digitado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com, no máximo, 150 palavras;

5. O artigo deve ser dividido em seções, numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado;

6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

7. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;

8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas sequencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções de trabalho;

9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos;

10. Não serão permitidos itens que identifiquem os autores do artigo dentro do texto, tais como: número de projetos de órgãos de fomento, endereço, *e-mail*, etc. Caso ocorra, a responsabilidade será inteiramente dos autores; e

11. No caso de o artigo ser aceito para a publicação após a avaliação dos pareceristas, serão encaminhadas as sugestões/comentários aos autores sem a sua identificação. Uma vez nesta condição, é de responsabilidade única dos autores fazer o *download* da formatação padrão da revista (em doc ou em LaTeX) para o envio da versão corrigida.