VIII ERMAC-R3

5º Encontro Regional de Matemática Aplicada e Computacional
20 a 22-Novembro-2008
Universidade Federal do Rio Grande do Norte - Natal/RN

Comparações entre os Modelos de Proporções e os Modelos Lineares Generalizados

Joseilme Fernandes Gouveia
Programa de Pós-graduação em Biometria e Estatística Aplicada, UFRPE,
52171-900, Recife, PE
E-mail: juniorfg99@hotmail.com,

Carlos Sergio Araújo dos Santos
Programa de Pós-graduação em Biometria e Estatística Aplicada, UFRPE,
52171-900, Recife, PE
E-mail: carlossergioaraujo@gmail.com.

Resumo

Diversos estudos recentes compreendem a análise de variáveis definidas no intervalo (0,1), como proporções. Os modelos mais utilizados na literatura são beta e simplex. Os modelos lineares generalizados (MLG) proposto por Nelder e Wedderburn (1973) possibilitam abrir um leque de opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a média e o preditor linear. Este trabalho consiste em comparar através de uma aplicação a dados reais o modelo proposto por Ferrari e Cribari-Neto (2004) e o desenvolvimento da regressão simplex com os MLGs. Define-se um resíduo para o modelo de regressão simplex, muito útil na análise de diagnóstico, a partir do trabalho de Espinheira, Ferrari e Cribari-Neto (2008). Além de apresentar uma forma geral das medidas de diagnósticos.

Palavras-chave

Regressão Beta, Regressão Simplex, Modelos Lineares Generalizados.

Introdução

Diversas situações compreendem a análise de variáveis de proporções. Pode-se citar, por exemplo, as proporções de casa que tem TV a cabo, de renda gasta com alimentações, etc. Muitos modelos são estudados para analisar variáveis contínuas definidas no intervalo (0,1), em função de outras variáveis, desde modelo de regressão linear a modelos com distribuição.

Kieschnick e McCullough (2003) separam esse tipo de variável em duas categorias. Os dados definidos no intervalo aberto (0,1), enquanto que na segunda, são definidos no intervalo [0,1] e seguem uma distribuição mista discreta-contínua. Aqui, vamos estudar variáveis que pertencem à primeira categoria, em específicos os modelos de regressão baseado na distribuição beta e distribuição simplex [6]. Além dos modelos lineares generalizados, que são uma extensão dos modelos normais lineares [4]. Faz-se uma comparação desses modelos, através de uma aplicação, dentre esses três modelos uma variação do critério de AKAIKE evidencia melhor adequação dos modelos lineares generalizados.

Métodos

0.1 Modelo de Regressão Beta

Ferrari e Cribari-Neto (2004) propuseram o modelo de regressão beta para situações em que a variável resposta é contínua e restrita ao intervalo (0,1) e está relacionada a outras variáveis através de uma estrutura de regressão. O modelo sugere uma reparametrização da distribuição beta, considerando-se a média da resposta e um parâmetro de dispersão.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, onde cada y_t, $t = 1, \ldots, n$, segue uma distribuição com média μ_t e parâmetro de dispersão desconhecido ϕ (constante para todo t). O modelo de regressão beta é definido pela distribuição beta e pelo componente sistemático

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

em que $\eta_t = x_t^T\beta$ é o preditor linear, $\beta = (\beta_1, \ldots, \beta_k)^T$ é um vetor de parâmetros desconhecidos a serem estimados $\left(\beta \in \Re^k\right)$, $x_t^T = (x_{t1}, \ldots, x_{tk})$ representa os valores de $k(k < n)$ variáveis explicativas que são assumidas fixas e conhecidas. A função $g(.)$ é chamada de função de ligação, uma função estritamente monótona e duplamente diferenciável que transforma valores do intervalo (0,1) em $\Re$.

Uma função de ligação utilizada é a logito, pois permite interpretação simples para os parâmetros de regressão. Em (1), segue que

$$g(\mu_t) = log\left(\frac{\mu_t}{1-\mu_t}\right) = x_t^T\beta, t = 1, \ldots, n.$$

Quando a função logito é usada, os parâmetros de regressão podem ser interpretados em termos de razão de chances (*odds ratio*).

0.2 Modelo de Regressão Simplex

Outra distribuição que pode ser utilizada para estudar uma variável resposta contínua e restrita ao intervalo (0,1) é a distribuição simplex [5]. A distribuição simplex faz parte de modelos de dispersão [1], que estendem os modelos lineares generalizados.

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes, onde cada y_t tem distribuição $S\left(\mu_t, \sigma^2\right)$, t=1,. . .,n. O modelo de regressão simplex é definido pela densidade da distribuição simplex, onde as médias μ_t são modeladas por

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

Em que $g(.)$ é uma função de ligação, estritamente monótona e duplamente diferenciável que transforma valores do intervalo (0,1) em $\Re$, $\beta = (\beta_1, \ldots, \beta_k)^T$ é o vetor dos parâmetros da regressão $\left(\beta \in \Re^k\right)$, $x_t^T = (x_{t1}, \ldots, x_{tk})$ são os valores conhecidos de k covariáveis e η_t é o preditor linear.

Defini-se um resíduo para o modelo de regressão simplex, muito útil na análise de diagnóstico, a partir do trabalho de Espinheira, Ferrari e Cribari-Neto (2008) por

$$r_t^{pp} = \frac{\hat{u_t}}{\sqrt{q_t(1-h_{tt})}},$$

em que $\hat{u_t} = -(1/2)d'(y_t; \hat{\mu}_t)$, h_{tt} é o t-ésimo elemento da diagonal de H e $q_t = \hat{Var}(u_t)$.

A matriz H é definda

$$H = \hat{A}^{1/2}X(X^T\hat{A}X)^{-1}X^T\hat{A}^{1/2}.$$

em que A é a matriz de pesos.

0.3 Modelos Lineares Generalizados

Nelder e Wedderburn (1972) mostraram que uma série de técnicas estatísticas, comumente estudadas separadamente, podem ser formuladas, de uma maneira unificada, como uma classe de modelos de regressão. A essa teoria unificadora de modelagem estatística, é uma extensão dos modelos clássicos de regressão, deram o nome de modelos lineares generalizados (MLG). Os MLGs consiste em abrir um leque de opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a média e o preditor linear.

A estimação neste estudo é realizada através do método da máxima verossimilhaça para todos os parâmetros, para as distribuições Normal, Binomial, Normal Inversa e Gama. Isto pode ser generalizado para outras distribuições.

Considerando n variáveis aleatórias independentes $y_1, \ldots, y_n$, cada uma com função densidade (ou probabilidade) na família exponencial da forma

$$f(y; \theta_i, \phi) = exp\left\{\phi\left[y\theta_i - b(\theta_i)\right] + c(y, \phi)\right\}, i = 1, \ldots, n \quad (2)$$

em que $b(.)$ e $c(.)$ são funções conhecidas. Para o modelo em (2) valem as seguintes relações:

$$E(y_i) = \mu_i = b'(\theta_i), Var(y_i) = \phi^{-1}V_i, i = 1, \ldots, n$$

sendo ϕ^{-1} o parâmetro de dispersão e $V = d\mu/d\theta$ a função da variância (carateriza a distribuição).

Os MLGs são definidos por (2) e pela componente sistemática

$$g(\mu_i) = \eta_i = X\beta, i = 1, \ldots, n$$

onde $g(.)$ é uma função de monótona e diferenciável, denominda função de ligação, e $X = (x_1, \ldots, x_n)$ é a matriz do modelo, $\beta = (\beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros e $\eta = (\eta_1, \ldots, \eta_n)^T$ é o preditor linear.

Algumas medidas de diagnóstico

Uma etapa importante do ajuste de um modelo de regressão e a análise de diagnóstico, pois nos permite verificar possíveis afastamentos das suposições feitas para o modelo e nos auxilia na identificação de observações extremas com alguma interferência desproporcional nos resultados do ajuste.

Um ponto alavanca possui um perfil diferente dos demais em relação aos valores das variáveis explicativas. Na prática, construímos um gráfico de h_{tt} versus o índice das observações t, $t = 1, \ldots, n$. Um alto valor de h_{tt} comparado com os valores das demais observações pode indicar que o ponto é de alavanca.

Um ponto aberrante é aquele que apresenta perfil diferente das demais obervações em relação aos valores da variável resposta, e também possui valor baixo na matriz de projeção H. Com isso, um ponto dificilmente é alavanca e aberrante.

Um ponto influente é aquele que exerce um peso desproporcional nas estimativas dos parâmetros do modelo, possui um perfil diferente dos demais no

que tange aos valores da variável resposta e apresenta valor alto na matriz de projeção H. Uma medida de influência de cada observação nas estimativas dos parâmetros é a distância de Cook [8]. Essa medida mede o afastamento entre a estimativa do vetor paramétrico utilizando todas as observações ($\hat{\beta}$) e sem a observação y_t ($\hat{\beta}_t$).

Uma medida utilizada para selecionar o modelo mais parcimonioso entre os ajustados, ou seja, que esteja bem ajustado e com um número reduzido de parâmetros, é o critério de informação de Akaike (AIC). O critério de informação de Akaike é dado por

$$AIC = -2l(\hat{\beta}) + 2k,$$

sendo $l(.)$ o logaritmo da função de verossimilhança atribuída aos dados, $\hat{\beta}$ o estimador de máxima verossimilhança com base no modelo ajustado e k a dimensão de $\hat{\beta}$.

O gráfico de probabilidade meio-normal com envelope simulado é uma ferramenta de diagnóstico muito útil para avaliarmos o ajuste do modelo. Este gráfico é construído baseado nos resíduos padronizados. Detalhes sobre sua construção podem ser encontrados em Neter, Kutner, Naschtheim Wasserman (1996). O envelope simulado é a banda de confiança. A ocorrência de pontos próximos ou fora da banda de confiança indicam que o modelo não está apropriado.

Resultados

Utilizamos os dados de oxidação de amônia, analisados originalmente por Brownlee (1965). Este estudo consiste em analisar a perda na conversão da amônia em ácido nítrico, em 21 dias de processos de produção de ácido nítrico em uma planta industrial. O ácido nítrico é utilizado na produção de fertilizantes, corantes, medicamentos, etc. O processo ocorre da seguinte forma: o gás amônia reage com o oxigêno do ar formando o óxido nítrico, que reage novamente com o oxigênio do ar produzindo o dióxido de nitrogênio. Este, por sua vez, reage com a água formando o ácido nítrico e também o óxido nítrico. Então, o óxido nítrico produzido é absorvido e reutilizado no processo. Há liberação de muito calor em todas essas reações químicas, sendo necessário o resfriamento do processo através da água.

A variável dependente (y) é a perda na conversão de amônia em ácido nítrico, que corresponde à proporção de amônia não convertida em ácido nítrico. O objetivo é modelar a proporção de amônia não convertida em ácido nítrico em função das covariáveis: corrente de ar (x_1), temperatura da água (x_2) utilizada no resfriamento do processo e a concentração de ácido nítrico (x_3).

Ajuste do Modelo de Regressão Beta

Para ajustar o modelo de regressão beta a esses dados, assumiremos que $y_1, \ldots, y_{21}$ são independentes e seguem uma distribuição beta com média $\mu_t = 1, \ldots, 21$, e o parâmetro de precisão ϕ desconhecido. Inicialmente, modelamos a média como

$$g(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_{t3} x_{t3}, \qquad (3)$$

sendo β_0 correspondente ao intercepto e $g(\mu_t)$ a função de ligação logito.

Os resultados do ajuste do modelo de regressão beta especificado em (3) são dados na tabela 1.

Tabela 1: Estimativas dos parâmetros do modelo de regressão beta ajustado para os dados de oxidação de amônia.

Parâmetros	Estimativa	Erro Padrão	p-valor
β_0	-7,7297	0,6795	0,0000
β_1	0,0294	0,0063	0,0000
β_2	0,0741	0,0189	0,0001
β_3	0,0029	0,0086	0,7390
ϕ	2361,3550	730,3160	

Aplicando as técnicas de diagnóstico, construindo os gráficos apresentados na Figura 1(a-d). A observação 2 aparece como possível ponto de alavanca e a observação 4 como possível ponto aberrante. A distância de Cook revela que as observacões 2, 4 e 21 são maiores que as demais observações, destacando-as como possíveis pontos influentes. A figura 1(d) apresenta o gráfico de probabilidade meio-normal. Notamos que os pontos encontram-se dentro da banda de confiança.

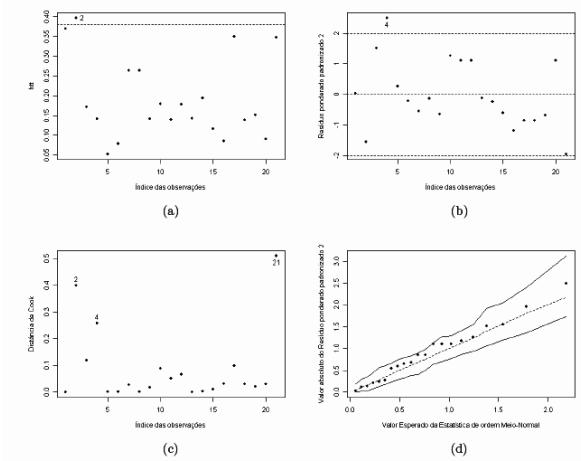

Figura 1: Gráficos de diagnóstico do modelo de regressão beta ajustado para os dados de oxidação de amônia.

Para avaliar se os pontos 2, 4 e 21 exercem alguma influência nas estimativas dos parâmetros,

ajustou-se novamente o modelo de regressão beta com a exclusão de cada um desses pontos, separadamente. Porém, não há convergência do processo de estimação dos parâmetros. Ajustamos, então, o modelo excluindo esses três pontos. Notamos que a exclusão das observacões (2, 4 e 21) altera bastante as estimativas dos parâmetros.

Realizando novas análises de diagnósticos para verificar se ainda há alguma observação atípica. O modelo de regressão beta sem as observações (1, 2, 4 e 21) é adequado aos dados de oxidação de amônia e a variável concentração de ácido nítrico não é signficativa para explicar a perda na conversão de amônia em ácido nítrico.

Tabela 2: Estimativas dos parâmetros do modelo de regressão beta ajustado para os dados de oxidação de amônia,sem as observações (1,2,4 e 21).

Parâmetros	Estimativa	Erro Padrão	p-valor
β_0	-8,0139	0,5869	0,0000
β_1	0,0529	0,0052	0,0000
β_2	0,0326	0,0129	0,0001
β_3	0,0000	0,0086	0,9820
ϕ	8734,8850	512,1235	

Ajuste do Modelo de Regressão Simplex

Consideremos que $y_1, \ldots, y_{21}$ são independentes e seguem uma distribuição simplex com média μ_t e parâmetro de dispersão σ^2 desconhecido. Utilizando a função de ligação logito, a média é modelada como

$$g(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_{t3} x_{t3}, \qquad (4)$$

em que β_0 corresponde ao intercepto.

Aplicando as técnicas de diagnósticos, observa-se há existência de pontos atípicos. Após sucessivas análises o modelo de regressão simplex sem as observações (4, 10, 16, 17, 20 e 21) foi o adequado. Em todos os ajustes do modelo de regressão simplex apressentados aqui, a variável concentração de ácido nítrico é não significativa.

Fazendo nova análise de diagnóstico, apresentadas na Figura 2, a observação 15 aparece como possível ponto influente. Não se destacam pontos de alavanca ou aberrante. No gráfico de probabilidade meio-normal, notamos que os pontos se encontram dentro da banda de confiança. Por fim, excluindo-se também a observação 15, verificamos que as estimativas se alteram pouco, mas as conclusões inferenciais são as mesmas.

Tabela 3: Estimativas dos parâmetros do modelo de regressão Simplex ajustado para os dados de oxidação de amônia, sem as observações (4, 10, 16, 17, 20 e 21).

Parâmetros	Estimativa	Erro Padrão	p-valor
β_0	-8,0448	0,6753	0,0000
β_1	0,0502	0,0042	0,0000
β_2	0,0263	0,0101	0,0000
β_3	0,0035	0,0089	0,9863
σ	0,4937	0,1254	

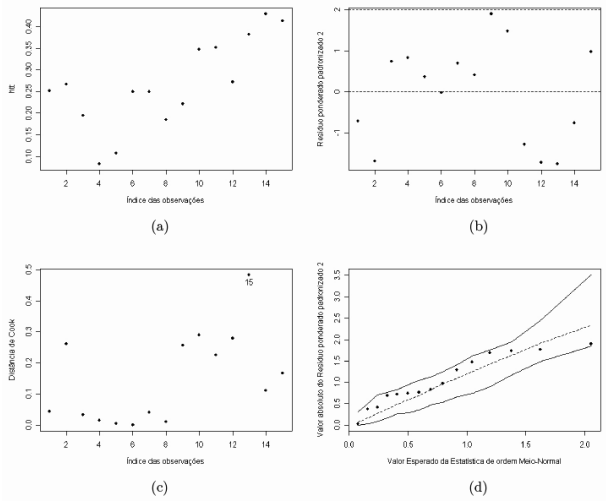

Figura 2: Gráficos de diagnóstico do modelo de regressão beta ajustado para os dados de oxidação de amônia, sem as observações (4, 10, 16, 17, 20 e 21).

Ajuste dos Modelos Lineares Generalizados

Aplicando o critério de AIC, temos o modelo selecionado foi gama com ligação identidade, cujas estimativas dos parâmetros seguem na Tabela 4. Este modelo apresentou desvio residual no valor de 0,17763 com 15 graus de liberdade. A inclusão do η^2 como uma variável explanatória não propriciou redução drástica no desvio, conclui-se que a função de ligação é satisfatória.

$$g(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_{t3} x_{t3}, \qquad (5)$$

Considerando uma distribuição gama com função de ligação identidade, o modelo (5) apresentou-se inicialmente alguns pontos atípicos. Após algumas análise de diagnósticos o modelo sem as observações (4, 21) adequou-se aos dados. Em todos os ajustes a variável ácido nítrico é não significativa.

Tabela 4: Estimativas dos parâmetros do modelo com distribuição gama com ligação identidade ajustado para os dados de oxidação de amônia, sem as observações (4, 21).

Parâmetros	Estimativa	Erro Padrão	p-valor
β_0	-0,0419	0,0049	0,0000
β_1	0,0078	0,0007	0,0000
β_2	0,0066	0,0020	0,0000
β_3	0,0231	0,0005	0,6869
ϕ	0,0118	0,0012	

A Figura 3(d) mostra o gráfico de probabilidade meio-normal. Observamos que os pontos estão dentro da banda de confiança, logo o modelo gama com

ligação identidade ajustado sem as observações (4, 21) é adequado para ajustar os dados.

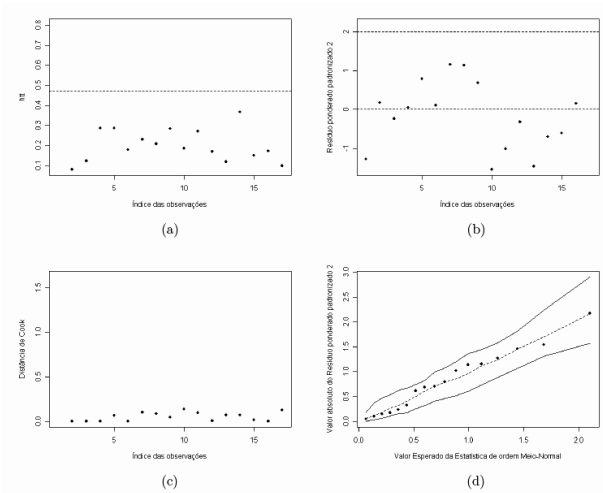

Figura 3: Gráficos de diagnóstico do modelo gama com ligação identidade ajustado para os dados de oxidação de amônia, sem as observações (4, 21).

.

Em forma geral, pode-se concluir que o modelo com distribuição gama e função de ligação identidade foi o que melhor ajustou-se aos dados de oxidação de amônia, pois seu ajuste excluiu menos observações e contém menor AIC em comparação com os modelos de regressão beta e simplex.

Conclusão

Apresentamos o modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004), o modelo de regressão simplex Barndorff-Nielsen e Jørgensen (1991) e os modelos lineares generalizados propostos por Nelder e Wedderburn (1972). Além da generalização das medidas de diagnósticos que permitem identificar possíveis ponto de alavanca, aberrante e/ou influentes a partir da matriz de projeção dos resíduos padronizados. Essas medidas puderam ser aplicadas tanto no caso do modelo de regressão beta, modelo de regressão simplex e nos modelos lineares generalizados.

Foi desenvolvida uma rotina computacional no R para ajuste do modelo de regressão simplex, baseada em algumas funções do pacote betareg. Implementamos também um programa para análise de diagnóstico baseado no resíduo ponderado padronizado 2 para os modelos. Os gráficos de probabilidade meio-normal com envelope simulado também foram construídos com esse resíduo.

Avaliou-se os modelos atráves das medidas de diagnósticos. Mostrando que apesar de existirem modelos específicos para o intervalo (0,1), os modelos lineares generalizados não devem ser desprezados. Pois, em devidas situações podem apresentar melhores ajustes, como nesse caso, pois o mesmo excluiu menos pontos e apresentou menor AIC para a obtenção de um bom ajuste.

Agradecimentos

A Conselho de Apoio a Pesquisa do Ensino Superior(CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro.

Referências

[1] B. Jørgensen, The theory of dispersion models, Chapman and Hall, London, 1997.

[2] H.Akaike, A new look at the statistical model identification. IEEE Trans. Auto Cntl AC-19, 6, 716723, 1974.

[3] J. Neter, H.N. Kutner, C.J. Naschtheim W. Wasserman, Appplied Linear Statistical Models, McGraw Hill, Chicago. 4.ed,(1996).

[4] J.A. Nelder e R.W.M. Wedderburn, Generalized linear models. Journal of the Royal Statistical Society A, local, V. 135, p. 370-384, (1972).

[5] K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, John Wiley and Sons, London. 2.ed, (1965).

[6] O.E. Barndorff-Nielsen B. Jørgensen, Some parametric models on the simplex, Journal of Multivariate Analysis 39, 106116,1991.

[7] P.L. Espinheira,S.L.P. Ferrari F. Cribari-Neto, On beta regression residuals, Journal of Applied Statistics, (2008).

[8] R.D. Cook, Detection of influential observations in linear regressions, Technometrics 19, 1518,(1997).

[9] R. Kieschnick B.D. McCullough, Regression analysis of variates observed on (0, 1): percentages, proportions and fraction, Statistical Modelling 3, 193213(2003).

[10] S.L.P Ferrari F. Cribari-Neto, Beta regression for modelling rates and proportions, Journal of Applied Statistics 31, 799815, (2004).