

Comparação entre Modelos Lineares Generalizados utilizando dados sobre AIDS na cidade de Recife

Kleber Napoleão Nunes de Oliveira Barros,

Programa de Pós-Graduação de Biometria e Estatística Aplicada, UFRPE

52171-900, Recife, PE

E-mail: napoleaok@hotmail.com,

Patrícia Silva Nascimento

Depto. de Estatística, UFPE,

17033-360, Recife, PE

E-mail: patriciaajp1408@hotmail.com.

Resumo

A AIDS é uma doença que se manifesta após a infecção do organismo pelo Vírus da Imunodeficiência Humana (HIV), identificado no Brasil em 1980. A qualidade de vida dos indivíduos portadores do vírus tem aumentado com o desenvolvimento de medicamentos cada vez mais eficazes. Este trabalho consiste em comparar Modelos Lineares Generalizados (MLGs) utilizando dados sobre AIDS na cidade de Recife (1984-2007) obtidos através do site do sistema DATASUS, organizados por faixa etária, raça e sexo. Nelder e Wedderburn (1972) introduziram os MLGs. A seleção de modelos é uma parte importante em modelagem estatística e envolve a procura de um modelo que descreva bem os dados. O MLG define-se pela distribuição de probabilidade para variável resposta, um conjunto de variáveis independentes e uma função de ligação. Uma importante característica é a suposição de independência, ou de não-correlação, entre as observações. Através dos MLGs, comparou-se o modelo Normal com função de ligação “log” e o modelo de Poisson com função de ligação “sqrt”. Com base no critério de informação de Akaike (AIC), concluiu-se que o segundo modelo se adequa aos dados. Fez-se posteriormente uma análise de resíduos, excluindo uma observação indicada como aberrante melhorando o modelo ajustado.

Palavras-chave

Modelos Lineares Generalizados, AIDS, Critério de Informação de Akaike.

Introdução

A aids foi identificada pela primeira vez no Brasil em 1980. Inicialmente, os segmentos populacionais afetados pela epidemia foram basicamente os homossexuais e bissexuais masculinos, os hemofílicos e demais pessoas que receberam sangue e hemoderivados. Em meados da década de 80, entre os casos de transmissão sangüínea, outro segmento populacional destacou-se entre os novos casos de aids, o de usuários de drogas injetáveis (UDI). Primeiramente, a epidemia manteve-se basicamente restrita a São Paulo e Rio de Janeiro, as maiores regiões

metropolitanas do País. Depois, observou-se a disseminação da epidemia para as diferentes regiões do Brasil. Apesar do progressivo registro de casos em todas as Unidades da Federação, ao longo das décadas de 80 e 90 e início da presente década, a epidemia da aids não se distribui de forma homogênea no País, observando-se uma maior concentração de casos nas regiões Sudeste e Sul, as mais desenvolvidas do País [1].

Nos últimos anos da década de 90, a frequência de casos entre mulheres cresceu consideravelmente, passando também a predominar a transmissão heterossexual como forma de difusão do HIV. De 1991 a 1999 as taxas de incidência passaram de 8,2 por 100.000 habitantes, em 1991, para 11,2, em 1999, e apresentam uma grande variação ao longo do território brasileiro. O número de novos casos apresenta-se estável desde 1996, com tendência de redução nos últimos dois anos devido à diminuição de casos em homens. Mesmo com os progressos alcançados na redução da mortalidade, o impacto da aids ainda é muito expressivo sobre a mortalidade em adultos na idade produtiva, representando a 4ª causa de óbito no grupo de 20 a 49 anos, no ano de 1996. A razão de óbitos por aids em adultos/adolescentes, que era de 28 por 1, em 1985, passou para 3 por 1, em 1995 [8].

No Brasil, 42% dos óbitos por AIDS ocorreram no grupo etário de 25-34 anos e 80% na faixa etária de 20-44 anos. De 1985 a 1995, a proporção de mortes secundárias à aids, em homens de 25 a 34 anos, aumentou de 0,2% para 11,3%; o sexo feminino (nessa mesma faixa etária) apresentou um incremento de 2,2% para 9,9%, neste mesmo período. As taxas de mortalidade para aids vêm decrescendo menos rapidamente entre as mulheres do que entre os homens. A diferença entre as tendências das taxas de mortalidade em homens e mulheres pode ser devida às diferenças de acesso aos cuidados para transmissão do HIV/Aids, entre outros fatores propriamente epidemiológicos, dessa forma expressando uma difusão mais vigorosa do HIV/Aids entre as mulheres em anos recentes [8]. Após 1996, pôde ser verificado um aumento de sobrevida e/ou redução do número de

VIII ERMAC-R3

8º Encontro Regional de Matemática Aplicada e Computacional

20 a 22-Novembro-2008

Universidade Federal do Rio Grande do Norte - Natal/RN

óbitos também entre as mulheres, ainda que obedecendo a uma tendência menos expressiva de declínio [4].

O objetivo deste artigo é comparar Modelos Lineares Generalizados utilizando dados sobre AIDS na cidade de Recife (1984-2007). Os dados foram obtidos pelo site do sistema DATASUS [3] e organizados de acordo com a Tabela 1:

Tabela 1: Descrição das variáveis em estudo

Variáveis	Descrição
y	Quantidade de casos de Aids
x_1	Faixa etária de 5 a 12 anos
x_2	Faixa etária de 13 a 19 anos
x_3	Faixa etária de 20 a 24 anos
x_4	Faixa etária de 25 a 29 anos
x_5	Faixa etária de 30 a 34 anos
x_6	Faixa etária de 35 a 39 anos
x_7	Faixa etária de 40 a 49 anos
x_8	Faixa etária de 50 a 59 anos
x_9	Faixa etária Maior que 60 anos
x_{10}	Raça/cor Branca
x_{11}	Raça/cor Preta
x_{12}	Raça/cor Amarela
x_{13}	Raça/cor Parda
x_{14}	Raça/cor Indígena
x_{15}	Sexo Masculino
x_{16}	Sexo Feminino

Metodologia

Durante muitos anos os modelos normais lineares foram utilizados para descrever a maioria dos fenômenos aleatórios. Com o desenvolvimento computacional ocorrido na década de 70, alguns modelos que exigiam a utilização de processos iterativos para a estimação dos parâmetros começaram a ser utilizados. O modelo normal não-linear, por exemplo, que assume uma estrutura não-linear para os parâmetros, teve um grande avanço com a proposta apresentada por Nelder e Wedderburn (1972), que propuseram os modelos lineares generalizados (MLGs) [6]. Este modelo contempla várias distribuições como, a binomial, Poisson, normal, gama, normal inversa, entre outras numa única classe, denominada família exponencial, generalizando o contexto clássico de regressão. Percebe-se que, tal família, consegue encaixar modelos com grande diversidade entre si, tanto no suporte, quanto em relação a natureza dos dados a serem analisados, ou mesmo a disposição gráfica modelada pelas diferentes distribuições. Eles evitam a aplicação de transformações para alcançar normalidade, linearidade e homogeneidade da variância. Dados de contagem ou dicotômicos, são exemplos que podem ser analisados via MLG [2].

Os MLGs podem ser utilizados quando se tem uma única variável aleatória Y associada a um conjunto de variáveis explicativas $x_1, \dots, x_p$. Para uma

amostra de n observações (y_i, x_i) em que $x_i = (x_{i1}, \dots, x_{ip})^T$ é o vetor de variáveis explicativas, o modelo envolve três componentes [7]: aleatório, sistemático e função de ligação.

- Componente aleatório: representado por um conjunto de variáveis aleatórias independentes $Y_1, \dots, Y_n$ provenientes de uma mesma distribuição que faz parte da família de distribuições com médias $E(Y_i) = \mu_i$. A fdp de Y é dada por:

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

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

- Componente sistemático (variáveis explicativas): as variáveis explicativas entram no modelo por uma estrutura linear, dada por:

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

em que $X = (x_1, \dots, x_n)^T$ é a matriz do modelo, $\beta = (\beta_1, \dots, \beta_p)^T$ é o vetor de parâmetro e η é o preditor linear.

- Função de ligação: função que relaciona o componente aleatório ao sistemático, da seguinte forma:

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

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

Para encontrar os estimadores dos parâmetros do modelo, utilizou-se o método iterativo Escore de Fisher, pelo método da máxima verossimilhança, o qual consiste em determinar o máximo absoluto do logaritmo da função verossimilhança. Ponto de alavanca é o que possui perfil diferente das demais observações em relação aos valores das variáveis explicativas. Para detectá-los, utiliza-se os elementos da diagonal principal de $H = \hat{P}^{1/2} X (X^T \hat{P} X)^{-1} X^T \hat{P}^{1/2}$, ou seja, h_{ii} , em que $\hat{P}^{1/2} z$ é o vetor resposta. Temos que y_i é ponto de alavanca, se $h_{ii} > 2p/n$. Ponto influente é aquele que exerce um peso desproporcional nas estimativas dos parâmetros do modelo, para detectá-los utiliza-se a Estatística Modificada de Cook (T_i). Se $T_i > 2(p/n)^{1/2}$, y_i será influente. Ponto aberrante (outlier) é aquele que apresenta perfil diferente dos demais em relação aos valores da variável resposta. Utiliza-se o resíduo padronizado versus o índice das observações, para detectá-los graficamente. O gráfico de probabilidade Meio-Normal (qq-plot) com envelope simulado (banda de confiança) é

construído baseado nos resíduos padronizados. A ocorrência de pontos próximos ou fora da banda de confiança indicam que o modelo não está apropriado. O Critério de Informação de Akaike (AIC) é uma medida utilizada para selecionar o modelo mais parcimonioso entre os ajustados, ou seja, que esteja bem ajustado e com um número reduzido de parâmetros [5], é dado por:

$$AIC = -2l(\hat{\beta}) + 2p \quad (4)$$

em que $l(\hat{\beta})$ é o logaritmo da função de verossimilhança dos dados e p é o número de parâmetros.

Para verificar a adequação da função de variância, utilizamos o método empírico do esboço dos resíduos versus o preditor linear η . Para que a função de variância seja adequada, este deve apresentar pontos dispersos ou simplesmente sem aparente tendência (linear, quadrática, logarítmica, entre outras). O teste da adequação da função de ligação e inclusão da covariável adicionada $\hat{\eta}^2$ consiste em se adicionar $u = \hat{\eta} \otimes \hat{\eta}$ como uma variável explanatória extra e examinar a mudança ocorrida no desvio, o que equivale ao teste da razão de verossimilhanças [2]. Se ocorrer uma diminuição, há evidência de que a função de ligação é insatisfatória e que se deve incluir u no modelo.

Resultados

Primeiramente, ajustou-se dois modelos: um com distribuição Normal e função de ligação “log”; e outro com distribuição Poisson e função de ligação “sqrt”. Comparando-se o AIC , escolheu-se o modelo com distribuição de Poisson com função de ligação “sqrt” pois obteve o menor AIC sendo 199,56. Retirou-se algumas variáveis que não foram significativas, faixas etárias (13-19, 20-24, Mais de 60) e raças (Amarela, Indígena) diminuindo ainda mais a estimativa do AIC para 197,5. Na Tabela 2 tem-se as estimativas dos coeficientes do modelo.

Observa-se na Tabela 2, que as variáveis faixa etária 5-12 (−0,4253), 25-29 (−0,1419), 30-34 (−0,1776), 35-39 (−0,2339), 40-49 (−0,1197), 50-59 (−0,2082) e a raça parda (−0,1310) influenciam negativamente. Sendo a faixa etária 5 a 12 anos a que mais influencia no modelo, ou seja, a cada novo caso de AIDS os indivíduos nessa faixa etária diminuem o total de casos em 42,53%. As variáveis raça branca (0,1434), preta (0,1976), e sexo masculino (0,2419), feminino (0,0935) influenciam positivamente. O

sexo masculino é a variável mais influente do modelo, isto é, a cada novo caso de AIDS os indivíduos do sexo masculino aumenta a estimativa de casos em 24,19%.

Tabela 2: Estimativa dos coeficientes para o modelo Poisson com função de ligação “sqrt”

Parâmetro	Estimativa	Erro padrão	P-valor
β_0	1,975	0,281	2,11e-12
β_1	−0,425	0,150	0,0045
β_4	−0,142	0,031	4,37e-06
β_5	−0,178	0,029	1,37e-09
β_6	−0,234	0,035	3,75e-11
β_7	−0,120	0,032	0,0002
β_8	−0,208	0,044	2,57e-06
β_{10}	0,143	0,050	0,0045
β_{11}	0,198	0,050	8,51e-05
β_{13}	−0,131	0,028	3,98e-06
β_{15}	0,242	0,025	< 2e-16
β_{16}	0,094	0,019	5,14e-07

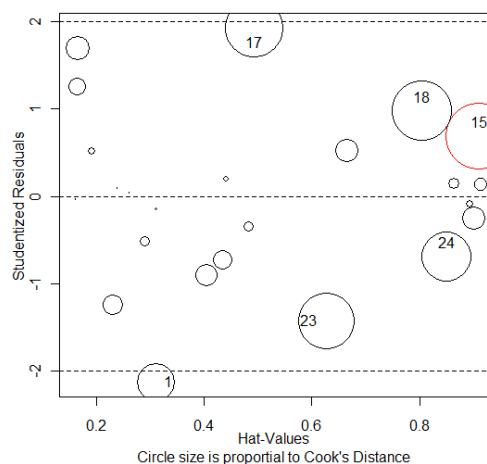


Figura 1: Gráfico de pontos de influência

Na Figura 1 verifica-se que a observação 15 (1998) é a mais influente enquanto os pontos 17, 18, 23 e 24 que correspondem, respectivamente, à 2000, 2001, 2006 e 2007 são de alavanca. Nota-se também que a observação 1 (1984) é um ponto aberrante, ou seja, o ano de 1984 apresenta perfil diferente dos demais em relação aos valores da variável resposta (número de casos de AIDS). A Figura 2 apresenta gráficos de diagnóstico que confirmam os resultados da Figura 1. Na figura 3 encontra-se um gráfico dos quantis da distribuições Poisson versus as observações com as bandas de confiança ajustada ao nível de significância de 5%. Observa-se que todos os pontos estão contidos no intervalo de confiança, indicando um bom ajuste do modelo. Nota-se ainda, que o ponto mais afastado da reta de ajuste é o ponto aberrante (1984).

VIII ERMAC-R3

8º Encontro Regional de Matemática Aplicada e Computacional

20 a 22-Novembro-2008

Universidade Federal do Rio Grande do Norte - Natal/RN

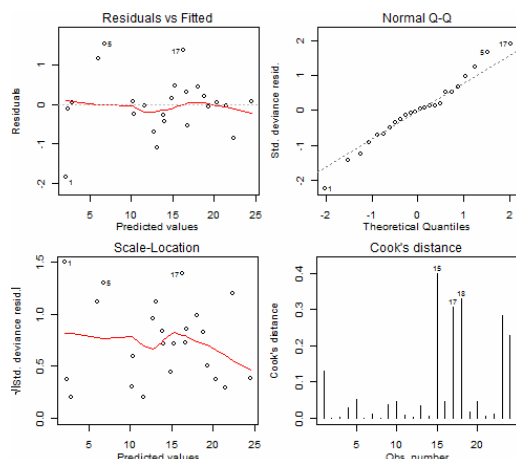


Figura 2: Gráficos de diagnóstico do modelo

Com base na análise dos resíduos, retirou-se o ponto aberrante (1984), e escolheu-se o modelo com distribuição de Poisson com função de ligação “sqrt”, em que verificou-se uma diminuição no AIC (190,93) indicando uma melhor adequação do modelo. Para verificar a adequação da função de variância utilizamos o método empírico do esboço dos resíduos versus o preditor linear η . Na Figura 4 aparecem os resíduos versus η (preditor linear). Como os pontos não parecem apresentar uma tendência, concluímos que a função de variância é adequada para o modelo. No teste da adequação da função de ligação, a inclusão de η^2 proporcionou uma diminuição do desvio residual de 8,135 para 1,061 e uma redução do AIC para 185,86, sugerindo que η^2 deve permanecer no modelo. Os valores estimados estão apresentados na Tabela 2.

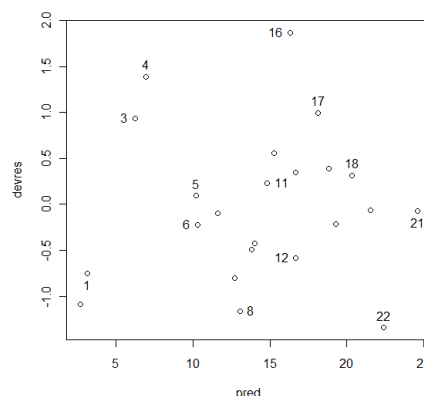


Figura 4: Gráfico de dispersão utilizado na verificação da adequação da função de variância.

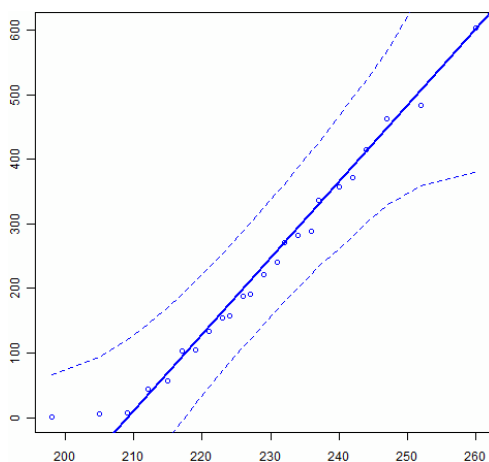


Figura 3: Qq-plot do modelo ajustado

Parâmetro	Estimativa	Erro padrão	P-valor
β_0	2,179	0,342	1,85e-10
β_1	-0,433	0,151	0,004166
β_4	-0,138	0,032	1,34e-05
β_5	-0,177	0,030	5,99e-09
β_6	-0,229	0,037	7,07e-10
β_7	-0,125	0,033	0,000139
β_8	-0,187	0,045	3,13e-05
β_{10}	0,150	0,052	0,004058
β_{11}	0,184	0,051	0,000268
β_{13}	-0,131	0,030	8,67e-06
β_{15}	0,274	0,032	< 2e-16
β_{16}	0,131	0,024	7,66e-08
β_{17}	-0,037	0,014	0,007841

Em que β_{17} representa o coeficiente estimado para a covariável η^2 .

VIII ERMAC-R3

8º Encontro Regional de Matemática Aplicada e Computacional

20 a 22-Novembro-2008

Universidade Federal do Rio Grande do Norte - Natal/RN

Na Tabela 3, observa-se que as variáveis faixa etária 5-12 ($-0,433$), 25-29 ($-0,138$), 30-34 ($-0,177$), 35-39 ($-0,229$), 40-49 ($-0,125$), 50-59 ($-0,187$), a raça parda ($-0,131$) e β_{17} ($-0,037$) influenciam negativamente. A faixa etária 5 a 12 anos continua sendo a mais influencia no modelo, ou seja, a cada novo caso de AIDS os indivíduos nessa faixa etária diminuem os casos em 43,3%. As variáveis raça branca (0,150), preta (0,184), e sexo masculino (0,274), feminino (0,131) influenciam positivamente. O sexo masculino também continua a variável mais influente do modelo, isto é, a cada novo caso de AIDS entre os homens aumentam 27,4% as estimativas da variável resposta.

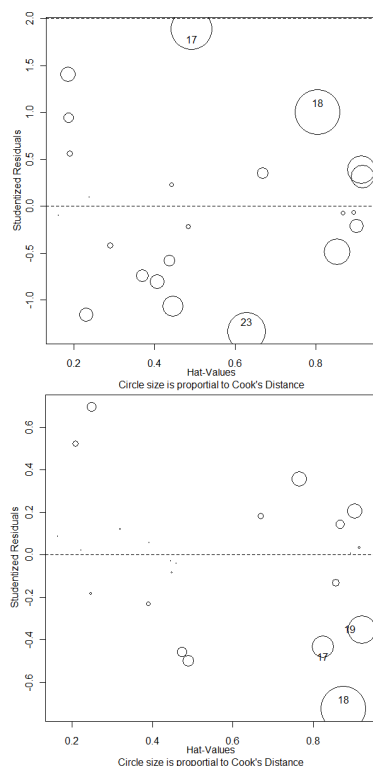


Figura 5: Gráfico de pontos de influência

Na Figura 5 verifica-se que o ponto 15 (1998) deixou de ser influente após a retirada da observação 1 (1984), ou seja, adequou-se ao modelo e as observações 17 (2000), 18 (2001) e 23 (2006) continuaram sendo de alavanca. Após a inclusão da covariável adicionada, notamos que a observação 23 deixou de ser ponto de alavanca e que as estimativas dos resíduos se aproximaram ainda mais do zero, não contendo pontos aberrantes, confirmando a melhora no ajuste. A Figura 6 confirma estes resultados. Na Figura 7, construiu-se um qqplot, tal como a Figura 3 com a exclusão da observação 1. Nota-

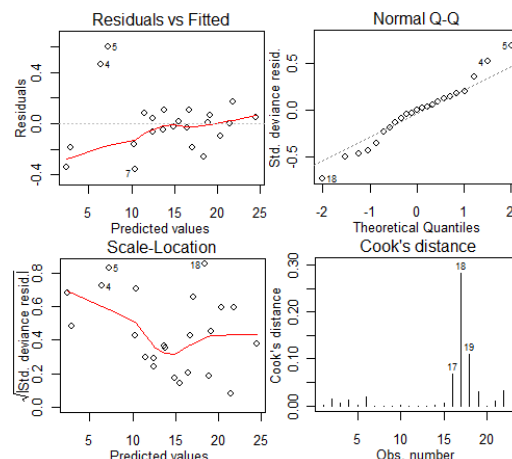


Figura 6: Gráficos de diagnóstico do modelo

se que, as observações ficaram, em média, mais próximas da reta de ajuste. Assim o modelo final é dado por:

$$\begin{aligned} \sqrt{\hat{\mu}} = & 2,17921 - 0,433x_1 - 0,138x_4 - 0,177x_5 \\ & - 0,229x_6 - 0,125x_7 - 0,187x_8 + 0,150x_{10} \\ & + 0,184x_{11} - 0,131x_{13} + 0,274x_{15} + 0,131x_{16} \\ & - 0,037\eta^2 \end{aligned}$$

em que:

Tabela 4: Descrição das variáveis do modelo final

Variáveis	Descrição
y	Quantidade de casos de Aids
x_1	Faixa etária de 5 a 12 anos
x_4	Faixa etária de 25 a 29 anos
x_5	Faixa etária de 30 a 34 anos
x_6	Faixa etária de 35 a 39 anos
x_7	Faixa etária de 40 a 49 anos
x_8	Faixa etária de 50 a 59 anos
x_{10}	Raça/cor Branca
x_{11}	Raça/cor Preta
x_{13}	Raça/cor Parda
x_{15}	Sexo Masculino
x_{16}	Sexo Feminino
η^2	Covariável adicionada

Com parâmetro de dispersão $\phi = 1$ e 10 graus de liberdade.

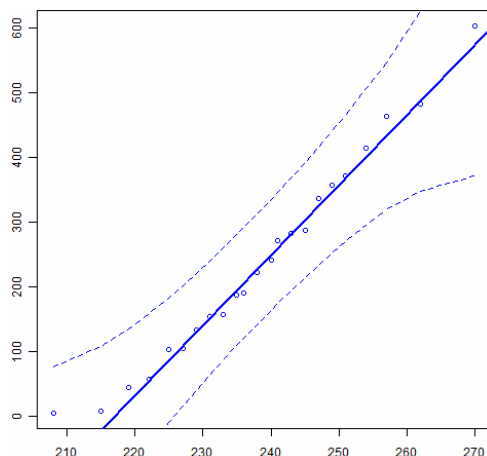


Figura 7: Qq-plot do modelo ajustado

Conclusões

Este artigo comparou Modelos Lineares Generalizados (MLGs) utilizando dados sobre AIDS na cidade de Recife (1984-2007), com dados obtidos pelo site do sistema DATASUS, organizados por faixa etária, raça e sexo. Através dos MLGs, comparou-se o modelo Normal com função de ligação “log” e o modelo de Poisson com função de ligação “log”. Com base no critério de informação de Akaike (AIC), concluiu-se que o segundo modelo se adequa melhor aos dados com $AIC = 197,5$. Com base na análise dos resíduos, observou-se uma observação aberrante (ano de 1984), excluiu-se esta observação, e escolheu-se o modelo com distribuição de Poisson com função de ligação “sqrt”, em que verificou-se uma diminuição no AIC (190,93) indicando uma melhor adequação do modelo. Fez-se um teste de adequação da função de variância e função de ligação, mostrando que os desvios diminuíram após a inclusão de uma nova covariável. Dessa forma, no modelo escolhido verificou-se que as variáveis faixas etárias de 5 a 12 anos e de 35 a 39 são responsáveis por um decréscimo de 43,3% e de 22,9% na variável resposta. Também notou-se que as variáveis sexo masculino e raça/cor negra influenciaram o modelo, respectivamente, com aumento de 0,274 e 0,184, isto é, a cada novo caso de AIDS entre indivíduos do sexo masculino ou de raça/cor negra aumentam o percentual de doentes em 27,4% e 18,4%.

Agradecimentos

A Universidade Federal Rural de Pernambuco (UFRPE) e a Fundação de Amparo à

Ciência e Tecnologia do Estado de Pernambuco (FACEPE) pelo apoio financeiro.

Referências

- [1] F. I. BASTOS, P. R. TELLES, E. A. CASTILHO, C. BARCELLOS, (1995). A epidemia de AIDS no Brasil. In: Minayo, MCS (org.). Os muitos brasis: Saúde e população na década de 80, pp. 245-268. , 2a edição 1999, São Paulo/Rio de Janeiro: Hucitec ABRASCO.
- [2] G. M. CORDEIRO, C. G. B. DEMÉTRIO, (2007). Modelos lineares generalizados. Minicurso para o 12º SEAGRO e a 52ª Reunião Anual da RBRAS, UFSM, Santa Maria, RS.
- [3] DATASUS - Departamento de Informática do SUS, [online], < [http : //www.datasus.gov.br](http://www.datasus.gov.br) >, acesso em 15 de agosto de 2008.
- [4] C. M. LOWNDES, F. I. BASTOS, K. M. GIFFIN, A. C. VAZ DOS REIS, E. D'ORSI, M. ALARY, (2000). Differential trends in mortality from AIDS in men and women in Brazil (1984-1995). AIDS 14(9):1269-73.
- [5] P. MCCULLAGH, J. A. NELDER, (1989). Generalized Linear Models, Chapman and Hall, London, 2.ed.
- [6] J. A. NELDER, R. W. M. WEDDERBURN, (1972). Generalized linear models. Journal of the Royal Statistical Society, A, 135, 370-384.
- [7] G. A. PAULA, (2004). Modelos de Regressão com Apoio Computacional. IME/USP, São Paulo.
- [8] G. SANTORO-LOPES, L. H. HARRISON, L. H. MOULTON, L. A. LIMA, A. M. DE PINHO, C. HOFER, M. SCHECHTER, (1998). Gender and survival after AIDS in Rio de Janeiro, Brazil. J Acquir Immune Defic Syndr Hum Retrovirol 19(4):403-7.