APLICAÇÃO DA ESTRATÉGIA DE ALOCAÇÃO NÃO-PROPORCIONAL PARA IDENTIFICAR *OUTLIERS* EM DADOS BINOMIAIS

TANIA MIRANDA NEPOMUCENA

2009

TANIA MIRANDA NEPOMUCENA

APLICAÇÃO DA ESTRATÉGIA DE ALOCAÇÃO NÃO-PROPORCIONAL PARA IDENTIFICAR *OUTLIERS* EM DADOS BINOMIAIS

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador

Prof. Marcelo Angelo Cirillo

LAVRAS
MINAS GERAIS – BRASIL
2009

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Nepomucena, Tania Miranda.
Aplicação da estratégia de alocação não proporcional para identificar *outliers* em dados binomiais / Tania Miranda Nepomucena. – Lavras : UFLA, 2009.
51 p. : il.

Dissertação (Mestrado) – Universidade Federal de Lavras, 2009.
Orientador: Marcelo Angelo Cirillo.
Bibliografia.

1. *Outliers*. 2. Estratégia não-proporcional. 3. Modelo binomial. 4. Funções de ligação. I. Universidade Federal de Lavras. II. Título.

CDD – 519.538

TANIA MIRANDA NEPOMUCENA

APLICAÇÃO DA ESTRATÉGIA DE ALOCAÇÃO NÃO-PROPORCIONAL PARA IDENTIFICAR *OUTLIERS* EM DADOS BINOMIAIS

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 18 de fevereiro de 2009.

Profª. Drª. Carla Regina Guimarães Brighenti	UFSJ
Prof. Dr. Renato Ribeiro de Lima	UFLA
Profª. Drª. Thelma Sáfadi	UFLA

Prof. Dr. Marcelo Angelo Cirillo
UFLA
(Orientador)

LAVRAS
MINAS GERAIS – BRASIL

Dedico a **DEUS**,

por me proporcionar esta importante conquista.

Ofereço

Aos meus pais,

Antonio Castro Nepomucena **e** ***Augusta Miranda Nepomucena***,

pelos ensinamentos de vida, esforço, amor e confiança.

Aos meus grandes amigos,

Zildelina **e** ***Alderino***,

pela constante proteção, bondade e incentivo.

Aos meus queridos irmãos,

Elias, Carmen, Luiz, Rubens, Silvia, Sandra, Sergio, Luzia e Cristiane, e as irmãs de coração ***Alzira, Luana e Amanda***,

pelo companheirismo, apoio e carinho.

AGRADECIMENTOS

A Deus, força maior de todo ser humano;

À minha querida mãe, pelo infinito amor, carinho e ensinamentos tão especiais;

Ao meu querido pai, pela coragem, pelo amor e confiança sempre dedicados a mim;

Ao professor Marcelo Angelo Cirillo, pela amizade, compreensão, orientação e apoio na elaboração deste trabalho;

À Universidade Federal de Lavras, pela oportunidade de concluir o mestrado em Estatística e Experimentação Agropecuária;

À Pró-Reitoria de Pesquisa e Pós-graduação da UFLA, pelo incentivo e apoio no desenvolvimento de pesquisas;

Ao Departamento de Ciências Exatas, pelo empenho e incentivo durante o curso de mestrado;

Ao Instituto Federal de Educação, Ciência e Tecnologia Baiano - Campus Catu-BA, pelo apoio irrestrito no decorrer da realização do curso;

Ao MEC - Secretaria de Educação Profissional e Tecnológica, por apoiar a capacitação de docentes;

À CAPES, pelo incentivo e apoio à qualificação de profissionais;

A todos os professores do Departamento de Ciências Exatas da UFLA, pela atenção, competência e colaboração na formação de novos profissionais;

Aos funcionários do Departamento de Ciências Exatas da UFLA, pela simpatia, dedicação e eficiência com que sempre nos atenderam;

Aos meus amigos e colegas do Instituto Federal de Educação, Ciência e Tecnologia Baiano - Campus Catu-BA, pelo incentivo, apoio e amizade;

Ao quarteto de primas queridas: Zildelina, Alzirinha, Luana, Amanda e meu grande amigo Alderino, pelo cuidado especial sempre dedicado a mim;

Aos meus grandes e queridos amigos Ana Paula, Augusto, Altemir, Edcarlos, Hiron, Paulo e Ricardo, pela constante amizade e compreensão;

Às companheiras Ana Patrícia, Elma e Layne, pela amizade e momentos saudáveis durante nosso convívio;

À minha amiga especial Ana Paula, pelas palavras de conforto e ombro amigo em todos os instantes;

À grande amiga Ana Patrícia, pela amizade que construímos, pelos momentos de risos e companheirismo;

Ao grande Augusto, amigo, companheiro e provocador de muitos risos;

Aos companheiros do curso Altemir, Augusto, Ana Patrícia, Ana Paula, Denise, Edcarlos, Hiron, Isabel, Paulo, Ricardo, Sthefania e Richardson, pela amizade, convivência, troca de conhecimentos e momentos de alegria;

A todos os colegas da pós-graduação em Estatística e Experimentação Agropecuária da UFLA;

Ao Devanil, pela preocupação com o nosso aprendizado nas aulas de Probabilidade e Inferência;

Aos meus queridos irmãos Elias, Carmen, Luiz, Rubens, Silvia, Sandra, Sergio, Luzia e Cristiane, pela união e compreensão;

A todos os familiares e demais pessoas que direta ou indiretamente contribuíram para esta minha conquista.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

NEPOMUCENA, Tania Miranda**. Aplicação da estratégia de alocação não-proporcional para identificar *outliers* em dados binomiais.** 2009. 51p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, MG.*

O método de análise robusta de variância que é apresentado, permite monitorar o efeito dos *outliers* por meio da estratégia não-proporcional. Para isso, utiliza-se a formação de subconjuntos, nos quais, as unidades amostrais são alocadas, baseando-se em apenas uma inspeção nos dados. Com o propósito de estender esse método para o modelo binomial, este trabalho propõe identificar *outliers* em dados de proporção com a aplicação da estratégia não-proporcional. Os dados utilizados são provenientes de um experimento realizado no Laboratório de Biologia do Departamento de Entomologia da ESALQ/USP. Foi implementado um algoritmo de uso no programa estatístico R para formação de subconjuntos previstos pelo método em estudo. Utilizando a função glm do programa R, estimou-se pelo método da máxima verossimilhança parâmetros do modelo binomial com as funções de ligação logit, probit e complemento log-log, também foram computados a deviance, os preditores lineares e os valores esperados dos modelos ajustados. A discussão em torno dos resultados foi realizada com base nos gráficos construídos. As estimativas dos parâmetros do modelo diferiram entre as funções de ligação. Para modelos diferentes, os preditores lineares apresentaram valores de magnitudes muito próximas. Observou-se que mesmo para funções de ligação distintas, os *outliers* podem ser monitorados. Isto exposto, concluiu-se que a metodologia proposta neste trabalho é viável e recomendável, podendo ser estendida para outros modelos pertencentes à classe de modelos lineares generalizados, sendo portanto, uma importante técnica de análise de dados para se detectar a presença de *outliers*.

---

*Orientador: Marcelo Angelo Cirillo – UFLA.

ABSTRACT

NEPOMUCENA, Tania Miranda. **Application of the strategy of non-proportional allocation to identify outliers in binomial data**. 2009. 51p. Dissertation (Master in Statistics and Agricultural Experimentation) – Federal University of Lavras, Lavras, MG.*

The method of robust analysis of variance that is displayed, allows monitoring the effect of outliers by the non-proportional strategy . To do this, using the formation of subsets, in which the sample units are allocated, based on one inspection to the data. With the aim of extending this method to the binomial model, this work proposes identify outliers in data of proportion with the application of the non-proportional strategy. The data used are from an experiment conducted in the Biology Laboratory, Department of Entomology of ESALQ / USP. An algorithm was implemented for use in the statistical program R to subsets provided by the method of study. Using the function glm in program R, was estimated by the method of maximum likelihood parameters of the binomial model with logit link-function, probit and complementary log-log, were also computed the deviance, the linear predictors and expected values of the adjusted models . The discussion around the results was based on graphs. Estimates of the parameters of the model differ between the link functions. For different models, the linear predictors had very similar values of magnitudes. It was observed that even for different link functions, the outliers can be monitored. This exposed, it was concluded that the methodology proposed in this paper is feasible and advisable, and may be extended to other models belonging to the class of generalized linear models, therefore, an important technique of data analysis to detect the outliers presence.

*Adviser: Marcelo Angelo Cirillo - UFLA.

1 INTRODUÇÃO

No processo de modelagem estatística ao se realizar uma análise exploratória do conjunto de dados, possivelmente a amostra poderá revelar a presença de *outliers*, que em síntese, trata-se de elementos que não obedecem a um padrão do conjunto de dados ao qual eles pertencem.

Os *outliers* requerem atenção especial, pois normalmente essas observações resultam em alguma violação das pressuposições necessárias para adequação ao modelo, produzindo conseqüentemente efeitos não confiáveis na eficiência dos estimadores (Tukey, 1960).

Algumas metodologias são encontradas na literatura relacionada à estimação robusta (Barnett, 1988; Atkinson & Riani, 2000), cujo interesse é estudar o comportamento de estimadores sob desvios das suposições paramétricas. Entretanto, convém ressaltar que o processo para obtenção de estimadores robustos torna-se um pouco complexo, pois exige do pesquisador conhecimentos mais específicos sobre teoria de estimação.

Dessa forma, tem-se a motivação para pesquisa de novas metodologias que sejam de melhor entendimento e que possam ser implementadas computacionalmente com maior facilidade.

Com o intuito de monitorar o efeito dos *outliers* sobre as estimativas de um modelo, Bertaccini & Varriale (2006) propuseram um método para detectar e investigar o efeito destas observações, considerando um modelo linear utilizado em conjunto com a técnica de análise de variância.

A metodologia proposta por esses autores permitiu não só analisar seus efeitos na estimação de parâmetros, mas também verificar o desempenho em testes de significância relacionados aos parâmetros de interesse.

Tendo por base essas informações, o presente trabalho tem por objetivo estender o método proposto por Bertaccini & Varriale (2006), aplicados em modelos lineares, para o modelo binomial utilizando as funções de ligação logit, probit e complemento log-log. Com essa finalidade, o método proposto para identificar *outliers* em dados binomiais é apresentado com uso da estratégia não-proporcional na modelagem de proporção de ovos parasitados de *A. Kuehniella* por fêmeas de *T. Galloi*, dos dados provenientes de um experimento realizado no Laboratório de Biologia do Departamento de Entomologia da ESALQ/USP.

2 REFERENCIAL TEÓRICO

2.1 Outliers na análise de variância

Um tópico muito discutido na inferência estatística é a presença de *outliers* em uma determinada amostra. Os *outliers* podem ser definidos como observações que aparentam ser inconsistentes com as demais pertencentes ao conjunto de dados (Barnett & Lewis, 1993). Anteriormente, Barnett (1988) expõe que tais observações provêm de distribuição não compatível com a distribuição que estrutura a amostra, ou são observações atípicas geradas por um modelo assumido.

Ainda se referindo aos *outliers*, Barnett & Lewis (1993) adotaram uma abordagem estatística para agrupar as causas de ocorrência de *outliers* durante a amostragem de dados:

- variedade inerente à população: os *outliers* são elementos que pertencem à população;
- erros de medição: ocorre na coleta dos dados. Pode ser causada por erros humanos, como a digitação de dados incorretos ou por erros de máquinas;
- erros de execução: ocorrem quando os dados são adquiridos através de amostragem de mais de uma população.

A média amostral é o melhor estimador não viesado da média populacional para dados de uma distribuição normal, mas mostra forte perda de eficiência no caso de contaminação dos dados (Bertaccini & Varriale, 2006).

A análise de variância é uma das técnicas estatísticas utilizada para testar diferença de médias entre populações normais. Na análise de variância, a

estatística F-Snedecor, comumente usada, é a razão dos desvios entre grupos e dentro dos grupos. Haja vista que a obtenção dos desvios envolve a média amostral, o valor de F-Snedecor é extremamente afetado quando calculado sob grupos que apresentam *outliers* (Bertaccini & Varriale, 2006).

Com o intuito de monitorar o efeito das observações *outliers* sobre as estimativas de um modelo linear, Bertaccini & Varriale (2006) propuseram um método para detectar e investigar o efeito destas observações sobre os resultados obtidos através da técnica de análise de variância em conjunto com o modelo linear.

O método consiste numa sequência de passos que através de uma inspeção dos dados pode alocar observações de modo não-proporcional.

A descrição do algoritmo utilizado para a alocação não-proporcional é a seguinte:

Etapa 1: Escolha do subconjunto inicial S_a

Determina-se um subconjunto S_a de unidades amostrais, selecionando em cada grupo i, a j-ésima observação y_{ij} que apresentar a menor diferença em módulo, de acordo com a seguinte condição:

$$\min \left| y_{ij} - \text{med}_i \right|, \; i = 1, ..., g \text{ e } j = 1, ..., r, \qquad (1)$$

sendo med_i a mediana amostral do i-ésimo grupo.

Etapa 2: Adicionando observações ao subconjunto S_a

Ajusta-se um modelo para o subconjunto S_a, em seguida calcula-se para as demais observações o quadrado do resíduo ordinário

$$e^2 = (y_{ij} - \hat{y}_{ij})^2 . \qquad (2)$$

A observação que apresentar o menor resíduo (2) é inserida ao subconjunto S_a. Na realização dessa etapa, obtém-se um novo subconjunto de observações, denominado por $S_{(a+1)}$, e o processo mais uma vez é repetido. O procedimento finaliza quando todas as observações são inseridas no modelo. É importante ressaltar que o índice a é obtido da igualdade $a = \sum_{i=1}^{g} a_i$, dado que os $a_i's$ são o número de observações do grupo i no passo a.

Etapa 3: Estimativas de parâmetros para identificação de *outliers*

Para cada subconjunto obtido em $S_{(a+1)}$, são computadas as estimativas das estatísticas de interesse do modelo.

Ao final do procedimento, os autores verificaram que é possível separar o grupo de *outliers* das demais observações através da análise gráfica feita para as estatísticas de interesse calculadas.

2.2 Modelos lineares generalizados

A classe de modelos lineares generalizados na verdade corresponde a uma extensão dos modelos lineares clássicos de Gauss-Markov, porém admite-se que a distribuição da variável resposta pertença à família exponencial e possibilite maior flexibilidade para ligação entre a média e a parte sistemática do modelo, além de não mais exigir as pressuposições básicas de normalidade, linearidade e homocedasticidade para a análise dos dados.

Outra particularidade interessante dos MLG é que a variância dos dados é modelada como uma relação conhecida da média (McCullagh & Nelder, 1989).

Segundo Dobson (2002), de um modo geral, a estrutura de um modelo linear generalizado resume-se nos seguintes itens:

- A variável dependente $\mathbf{Y}_i$ apresenta distribuição pertencente à família exponencial;
- As variáveis explanatórias x_{i1}, x_{i2}, **...,** x_{ip} e um vetor de parâmetros $\boldsymbol{\beta}$;

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- Uma função de ligação $g(.)$ de modo que $\eta_i = g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, sendo η_i o preditor linear e $\mu_i = E(Y_i)$.

2.3 Família exponencial

Uma variável aleatória Y com uma distribuição de probabilidade de parâmetro θ pertence à família exponencial se pode ser escrita conforme a seguinte parametrização (Dobson, 2002):

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \qquad \text{ou}$$

$$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right], \tag{3}$$

onde a, b, s e t são funções conhecidas, sendo $s(y) = \exp\left[d(y)\right]$ e $s(y) = \exp\left[d(y)\right]$.

A família exponencial na forma canônica é definida a partir de (3), considerando que as funções $a(y) = y$ e $b(\theta) = \theta$, ou seja, são iguais à função identidade. A presença de outros parâmetros diferentes de θ, podem ser considerados como parâmetros de perturbação, formando partes das funções a, b, c e d.

2.4 Distribução binomial

Em um grande número de circunstâncias é comum a ocorrência de observações que admitem apenas dois resultados, comumente conhecidos por "sucesso" ou "fracasso", considerando-se "sucesso" o evento de interesse. Entre várias situações, é possível citar como exemplos, o resultado do diagnóstico de um exame laboratorial, positivo ou negativo, e resultado de teste de aptidão aplicado a um estudante, aprovado ou reprovado.

Uma variável aleatória X pode ser definida como igual a 1, se ocorre "sucesso", ou igual a zero se ocorre "fracasso", de tal forma que as probabilidades atribuídas correspondem a

$$\mathrm{P}\left(\mathrm{X}=1\right)=\pi \quad \text{e} \quad \mathrm{P}\left(\mathrm{X}=0\right)=1-\pi \ , \tag{4}$$

e são conhecidas como variáveis aleatórias de Bernoulli.

Uuma variável aleatória Y, número de sucessos em amostras de tamanho m, é definida ter distribuição binomial se sua função de densidade discreta é dada por

$$\mathrm{P}\left(\mathrm{Y}=\mathrm{y}\right)=\binom{m}{\mathrm{y}}\pi^{\mathrm{y}}\left(1-\pi\right)^{m-\mathrm{y}},\ \mathrm{y}=0,1,...,m\,, \tag{5}$$

sendo os m ensaios independentes, com probabilidade de sucesso π constante (Mood et al., 1974) .

A distribuição binomial tem notação da forma $\mathrm{Y}\sim\mathrm{Bin}\left(m,\pi\right)$, sendo que o parâmetro m corresponde ao número de realizações de ensaios de Bernoulli e o parâmetro π a probabilidade do evento de interesse ocorrer.

Assim, a distribuição de Bernoulli é um caso particular da distribuição binomial quando $m=1$.

Então, pode-se afirmar que a distribuição binomial corresponde à soma $\mathrm{Y}=\sum_{\mathrm{i}=1}^{m}\mathrm{X}_{\mathrm{i}}$ de m variáveis aleatórias $\mathrm{X}_1,...,\mathrm{X}_m$ de Bernoulli, independentes e identicamente distribuídas.

O valor esperado e a variância da distribuição de probabilidade binomial são dados respectivamente por

$$\mathrm{E}(\mathrm{Y}) = m\pi \quad \text{e} \quad \mathrm{Var}(\mathrm{Y}) = m\pi(1-\pi). \tag{6}$$

2.5 Representação do modelo binomial na família exponencial

De acordo com a seção (2.3), uma distribuição pertence à família exponencial se é possível representá-la da forma (3). Ao se considerar a distribuição binomial $\mathrm{Bin}(m,\pi)$ cuja função de probabilidade corresponde a

$$\begin{aligned}
\mathrm{f}(\mathrm{y};\pi) &= \binom{m}{\mathrm{y}} \pi^{\mathrm{y}} (1-\pi)^{m-\mathrm{y}} \\
&= \exp\left\{ \log\left[\binom{m}{\mathrm{y}} \right] + \mathrm{y}\log(\pi) + (m-\mathrm{y})\log(1-\pi) \right\} \\
&= \exp\left\{ \log\left[\binom{m}{\mathrm{y}} \right] + \mathrm{y}\log(\pi) + m\log(1-\pi) - \mathrm{y}\log(1-\pi) \right\} \\
&= \exp\left\{ \mathrm{y}\log\left(\frac{\pi}{1-\pi} \right) + m\log(1-\pi) + \log\left[\binom{m}{\mathrm{y}} \right] \right\},
\end{aligned} \tag{7}$$

sendo portanto, um membro da família exponencial com as funções determinadas por $\mathrm{a}(\mathrm{y}) = \mathrm{y}$, $\mathrm{b}(\pi) = \log\left(\frac{\pi}{1-\pi} \right)$, $\mathrm{c}(\pi) = m\log(1-\pi)$ e

$d(y) = \log\binom{m}{y}$, podendo-se dizer que $b(\pi)$ é o parâmetro natural ou canônico[1].

2.6 Funções de ligação

Inicialmente na seção (2.2) foi dito que uma função de ligação $g(.)$ relaciona o preditor linear η à média μ do vetor de dados $\mathbf{y}$. É uma função que possui inversa e é diferenciável.

Um dos componentes de extrema importância na especificação dos MLG é a escolha da função de ligação, visto que a escolha adequada não só simplifica na obtenção das estimativas de máxima verossimilhança dos parâmetros do modelo, mas também, no cálculo do intervalo de confiança para a média da resposta (Myers et al., 2002).

A utilização de uma ou outra função de ligação, e conseqüentemente, a escolha do modelo de regressão a utilizar, dependem da situação considerada. Em geral, a adaptabilidade dos modelos probit e logit é bastante semelhante, já que as funções correspondentes não se afastam muito uma da outra após um ajuste adequado dos correspondentes preditores lineares.

No modelo normal linear, a média e o preditor linear são idênticos e ambos podem assumir qualquer valor pertencente ao conjunto dos números reais. Em modelos que assumem a distribuição binomial, existe a restrição de que $0<\pi<1$, portanto o domínio da função de ligação é o intervalo real $(0, 1)$, enquanto seu contradomínio é a reta real.

[1] Neste trabalho é considerado ln(.)= log(.).

Para assegurar a restrição que π é uma probabilidade e, portanto pertence ao intervalo $(0,\ 1)$, a modelagem frequentemente é realizada com uso de uma distribuição de probabilidade acumulada

$$\pi = \int_{-\infty}^{x} f(u)du\,. \tag{8}$$

A função densidade de probabilidade $f(u)$ é chamada distribuição de tolerância (Dobson, 2002).

A Figura 2 mostra as distribuições de tolerância cujas funções são não-lineares, correspondentes aos modelos logit, probit e complemento log-log. A ideia é fazer uma transformação capaz de linearizar a curva (Cordeiro & Demétrio, 2007).

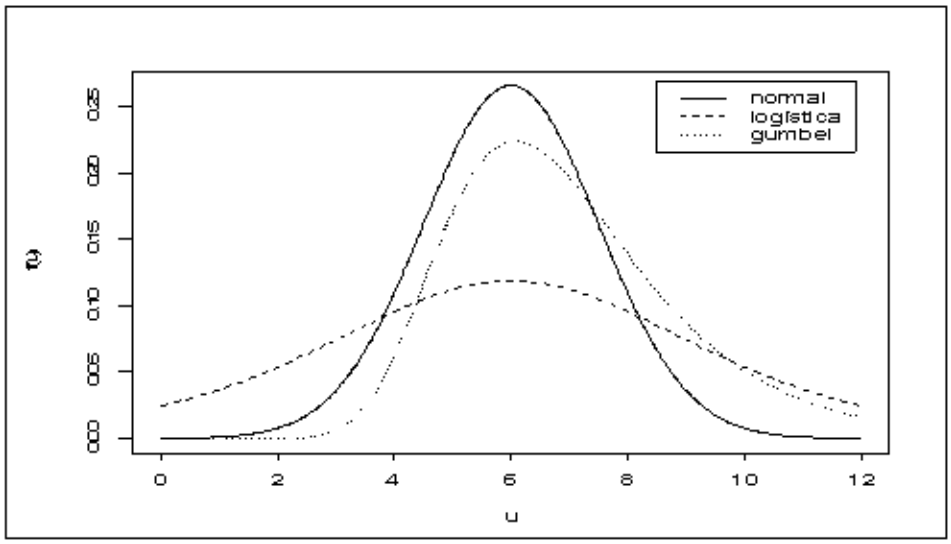


FIGURA 1 Distribuições de tolerância para o modelo logit, probit e complemento log-log.

A distribuição de tolerância logística definida por

$$f(u) = \frac{\beta_2 \exp(\beta_1 + \beta_2 u)}{\left[1 + \exp(\beta_1 + \beta_2 u)\right]^2},$$

modela a proporção π por meio da integração

$$\pi = \int_{-\infty}^{x} f(u)du = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)},$$

que é linearizada pela expressão

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x, \qquad (9)$$

denominada função de ligação **logit**.

Com a distribuição normal, pode-se escrever que

$$\pi = \frac{1}{\sigma\sqrt{2\pi^*}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] du$$
$$= \Phi\left(\frac{x-\mu}{\sigma}\right),$$

sendo $\pi^* = 3{,}1415...$ e Φ denota a função de probabilidade acumulada da distribuição normal padrão $N(0,1)$, linearizada pela função inversa

$$\Phi^{-1}(\pi) = \beta_1 + \beta_2 x, \qquad (10)$$

dado $\beta_1 = -\frac{\mu}{\sigma}$ e $\beta_2 = \frac{1}{\sigma}$, conhecida por função de ligação **probit**.

Usada como distribuição de tolerância, a Gumbel ou valor extremo

$$f(u) = \beta_2 \exp\left[\left(\beta_1 + \beta_2 u\right) - \exp\left(\beta_1 + \beta_2 u\right)\right], \text{ fornece}$$

$$\pi = \int_{-\infty}^{x} f(u)du = 1 - \exp\left[-\exp\left(\beta_1 + \beta_2 x\right)\right],$$

que é linearizada pela função de ligação **complemento log-log**:

$$\log\left[-\log\left(1-\pi\right)\right] = \beta_1 + \beta_2 x\,. \qquad (11)$$

Em resumo, as funções de ligação com as respectivas distribuições acumuladas são descritas na tabela abaixo:

TABELA 1 Funções de ligação com as respectivas funções de tolerância.

Nome	**função de ligação**	**distribuição**	**função de distribuição**
Logit	$g_1(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$	logística	$f(u) = \frac{\beta_2 \exp\left(\beta_1 + \beta_2 u\right)}{\left[1 + \exp\left(\beta_1 + \beta_2 u\right)\right]^2}$
Probit	$g_2(\pi) = \Phi^{-1}(\pi)$	normal	$f(u) = \frac{1}{\sigma\sqrt{2\pi^*}} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right]$
complemento log-log	$g_3(\pi) = \log\left[-\log\left(1-\pi\right)\right]$	gumbel	$f(u) = \beta_2 \exp\left[\left(\beta_1 + \beta_2 u\right) - \exp\left(\beta_1 + \beta_2 u\right)\right]$

McCullagh & Nelder (1989) ao compararem as funções logit, probit e complemento log-log (Figura 2), verificaram que estas apresentam comportamento aproximadamente linear quando π pertence ao intervalo $(0, 1)$ e, por esta razão, fica mais difícil concluir com qual dessas funções se tem um melhor ajuste.

As funções logit e probit são simétricas em torno de $0,5$. Para pequenos valores de π, as ligações logit e complemento log-log encontram-se bastante próximas, decaindo mais rapidamente que a probit. No entanto, quando π se aproxima de 1, a complemento log-log cresce mais lentamente do que as ligações probit e logit.

A importância de se estudar o comportamento das funções de ligação em relação ao preditor linear tem sido justificada por alguns autores. Desenvolveram estudo com a finalidade de verificar o efeito da especificação inadequada da função de ligação em modelos lineares generalizados com distribuição binomial. Os modelos ajustados com a função de ligação canônica, considerada correta, foram comparados com outros modelos de ajustes realizados com funções de ligação não canônicas. Foi constatado que as funções de ligação não canônicas causaram maior impacto nas inferências dos resultados, gerando intervalos de confiança que reduziram as taxas de cobertura, à medida que aumentavam o tamanho das amostras (Myers et al., 2002).

Andrade (2007) comparou os ajustes obtidos com função de ligação correta, usada para gerar dados, com uma função de ligação dita incorreta. De um modo geral, quando a função de ligação correta foi a logit ou probit, os ajustes trouxeram resultados satisfatórios e semelhantes para as estimativas das médias.

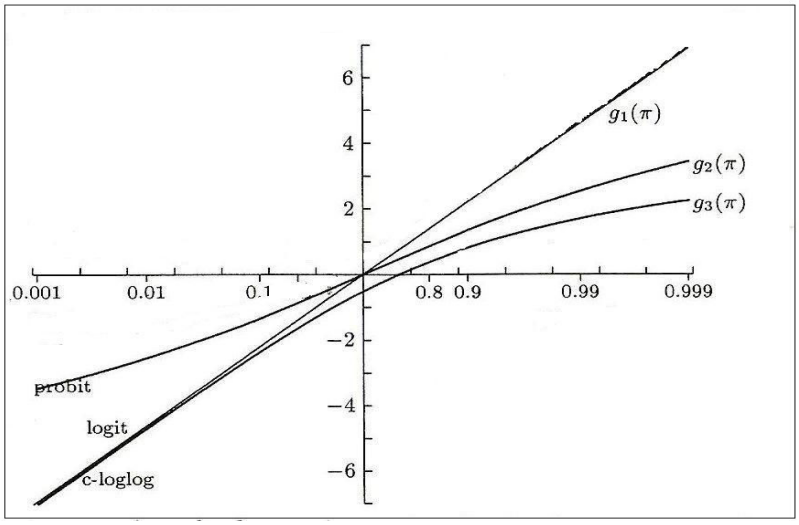


FIGURA 2 Funções de ligação logit $g_1(\pi)$, probit $g_2(\pi)$ e complemento log-log $g_3(\pi)$.

Para Andrade (2007) a má especificação da função de ligação afetou os resultados das médias estimadas quando a função de ligação correta é a complemento log-log e o modelo ajustado considerou a função de ligação logit, neste caso, as médias em torno de 0.5 e próximo de 1 ficaram abaixo do esperado, isto observado mais claramente para as amostras de maior tamanho.

No caso em que a utilização da função logit (incorreta) não trouxe bons resultados em relação aos intervalos de confiança para as médias, foi feita análise dos resíduos das observações da variável resposta e não foi detectada a especificação incorreta em nenhum cenário, os vieses relativos aos estimadores de máxima verossimilhança para as médias, indicaram que apesar das taxas de coberturas serem baixas para alguns ajustes utilizando a função de ligação logit (incorreta), as estimativas das médias estavam próximas aos seus verdadeiros valores, pois os vieses observados foram pequenos.

2.7 Estimação dos parâmetros do modelo binomial logit via método de máxima verossimilhança

Na família exponencial, tendo por base a representação do modelo binomial descrito na parametrização (3), nesta seção descreve-se a estimação do parâmetro de um modelo binomial, utilizando a função de ligação logit. Com esse propósito, é considerado o caso mais geral, assumindo algumas suposições.

Sejam M variáveis aleatórias independentes, isto é, $Y_1, Y_2, ..., Y_M$ de modo que cada y_i representa o número de sucessos em M diferentes subgrupos, conforme ilustra a Tabela 2.

TABELA 2 Freqüências de distribuições $\text{Bin}(m_i, \pi_i)$, $i = 1, 2, ..., M$ com variável aleatória Y_i.

	Subgrupos			
	1	2	...	M
Sucessos	y_1	y_2	...	y_M
Fracassos	$m_1 - y_1$	$m_2 - y_2$	...	$m_M - y_M$
Totais	m_1	m_2	...	m_M

Dado que cada Y_i pertence à família exponencial, então a função de verossimilhança é definida por

$$L(\pi_i; y_i) = \exp\left[\sum_{i=1}^{M} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{M} m_i \log(1-\pi_i) + \sum_{i=1}^{M} \log\binom{m_i}{y_i}\right]. \qquad (12)$$

Calculando o log da função de verossimilhança temos:

$$\log\left[L\left(\pi_i ; y_i\right)\right] = \sum_{i=1}^{M} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{M} m_i \log\left(1-\pi_i\right) + \sum_{i=1}^{M} \log\binom{m_i}{y_i}$$

$$\ell\left(\pi_i ; y_i\right) = \sum_{i=1}^{M}\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log\left(1-\pi_i\right) + \log\binom{m_i}{y_i}\right]. \qquad (13)$$

Em (13) pode-se observar que o parâmetro canônico[2] é dado por $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$. Utilizando-se da igualdade (9) é mostrado que

$$\begin{aligned}
&\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i \\
&\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_1 + \beta_2 x_i\right) \\
&\pi_i = \exp\left(\beta_1 + \beta_2 x_i\right) - \pi_i\left[\exp\left(\beta_1 + \beta_2 x_i\right)\right] \\
&\pi_i\left[1 + \exp\left(\beta_1 + \beta_2 x_i\right)\right] = \exp\left(\beta_1 + \beta_2 x_i\right) \\
&\pi_i = \frac{\exp\left(\beta_1 + \beta_2 x_i\right)}{1 + \exp\left(\beta_1 + \beta_2 x_i\right)} \\
&\pi_i = \frac{\exp\left(\theta_i\right)}{1 + \exp\left(\theta_i\right)} \qquad (14)
\end{aligned}$$

Da expressão (14) é verificado que $\theta_i = \beta_1 + \beta_2 x_i$. Fazendo a substituição em (13), resulta na expressão:

[2] Segundo Cordeiro & Lima Neto (2004), o parâmetro canônico caracteriza a distribuição de probabilidade membro da família exponencial.

$$\ell(\pi_i; y_i) = \sum_{i=1}^{M} \left\{ y_i(\beta_1 + \beta_2 x_i) - m_i \log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right] + \log\left[\binom{m_i}{y_i}\right] \right\} \qquad (15)$$

Outra relação que pode ser enunciada é dada por

$$\eta = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_i . \qquad (16)$$

Dobson (2002) cita que a derivada de $\ell(\pi_i; y_i)$ com relação ao β_p é chamada de estatística escore e resulta nas seguintes funções $U_p's$ dependentes de y_i

$$U_1 = \frac{\partial \ell(\pi_i; y_i)}{\partial \beta_1} = \sum_{i=1}^{M} \left\{ y_i - m_i \left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right] \right\}$$
$$= \sum_{i=1}^{M} (y_i - m_i \pi_i). \qquad (17)$$

$$U_2 = \frac{\partial \ell(\pi_i; y_i)}{\partial \beta_2} = \sum_{i=1}^{M} \left\{ y_i x_i - m_i \left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right] \right\}$$
$$= \sum_{i=1}^{M} x_i (y_i - m_i \pi_i). \qquad (18)$$

O valor esperado das derivadas parciais de segunda ordem dos escores $U_p's$ fornece a matriz de informação de Fisher $\mathcal{I}$, mencionada por McCullagh & Nelder (1989). Os elementos ι_{pq} que a compõe são dados por

$$\iota_{pq} = E\left[\frac{-\partial^2 \ell(\pi_i; y_i)}{\partial\beta_p \partial\beta_q}\right] = E\left[\frac{\partial \ell(\pi_i; y_i)}{\partial\beta_p}\frac{\partial \ell(\pi_i; y_i)}{\partial\beta_q}\right] = E\left[U_p U_q\right]. \quad (19)$$

Portanto, apropriando-se dos resultados (17) e (18), a matriz de informação de Fisher $\mathcal{I}$ fica definida pelos termos

$$\mathcal{I} = \begin{pmatrix} \sum m_i \pi_i (1-\pi_i) & \sum m_i x_i \pi_i (1-\pi_i) \\ \sum m_i x_i \pi_i (1-\pi_i) & \sum m_i x_i^2 \pi_i (1-\pi_i) \end{pmatrix}. \quad (20)$$

Para a estimação de máxima verossimilhança usando o método de escore, uma alternativa é aplicação do processo iterativo dado por

$$\mathcal{I}^{(k-1)}\mathbf{b}^{(k)} = \mathcal{I}^{(k-1)}\mathbf{b}^{(k-1)} + U_p^{(k-1)}, \quad (21)$$

sendo que $\mathbf{b}^k$ são estimativas do vetor de parâmetros $\boldsymbol{\beta}$ na k-ésima iteração.

Através de rearranjo dos termos (McCullagh & Nelder, 1989), a equação (21) pode ser reescrita como

$$\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X}\boldsymbol{\beta}^{(k)} = \mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{Z}^{(k-1)}, \quad (22)$$

sendo as matrizes $\mathbf{Z}$ e $\mathbf{W}$ estimadas da seguinte forma:

- mantendo a função de ligação em termos da média $\mu_i = m_i \pi_i \Rightarrow \pi_i = \frac{\mu_i}{m_i}$, obtém-se

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\mu_i}{m_i - \mu_i}\right) = \log(\mu_i) - \log(m_i - \mu_i). \qquad (23)$$

- calculando a derivada parcial de η_i em relação à média μ_i:

$$\begin{aligned}\frac{\partial \eta_i}{\partial \mu_i} &= \frac{1}{\mu_i} + \frac{1}{m_i - \mu_i} \\ &= \frac{m_i - \mu_i + \mu_i}{\mu_i (m_i - \mu_i)} \\ &= \frac{m_i}{m_i \pi_i (m_i - m_i \pi_i)} \\ &= \frac{1}{m_i \pi_i (1-\pi_i)}. \end{aligned} \qquad (24)$$

A matriz diagonal $\mathbf{W}$ é definida de forma que os elementos w_{ii}'s são obtidos por

$$\frac{1}{\mathrm{var}\left(Y_i\right)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{\mathrm{var}\left(Y_i\right)}\left\{\frac{\partial\left[\frac{m_i \exp\left(\eta_i\right)}{1+\exp\left(\eta_i\right)}\right]}{\partial \eta_i}\right\}^2$$

$$= \frac{1}{m_i\left[\pi_i\left(1-\pi_i\right)\right]}\left[m_i \pi_i\left(1-\pi_i\right)\right]^2$$

$$= m_i\left[\pi_i\left(1-\pi_i\right)\right] = w_{ii}. \qquad (25)$$

Em relação à matriz $\mathbf{Z}$, cada elemento z_i é definido pela seguinte expressão

$$z_i = \eta_i + \left(y_i - \mu_i\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) = \eta_i + \frac{\left[y_i - m_i \pi_i\right]}{m_i \pi_i\left(1-\pi_i\right)}. \qquad (26)$$

Cabe ressaltar que os programas computacionais de ajustamento do modelo linear generalizado sempre utilizam o método escore de Fisher para calcular as estimativas dos $\theta's$, isso devido à maior probabilidade de o algoritmo de Newton-Raphson não convergir.

2.8 Deviance e o teste da razão de verossimilhanças

A deviance se resume numa medida de discrepância, na escala do logaritmo da função de verossimilhança, dos valores ajustados $\hat{\pi}'s$ em relação aos dados observados $y's$. É uma estatística facilmente computada para qualquer modelo linear generalizado (Cordeiro & Demétrio, 2007).

Neste trabalho, o cálculo da deviance para o modelo binomial é ilustrado considerando Ω como um modelo completo, pelo fato do mesmo apresentar um

parâmetro para cada observação. Assim, pode-se dizer que o modelo é ajustado perfeitamente aos dados, visto que as estimativas das médias são iguais às observações. Nesse modelo, assume-se que toda a variação é devida à componente sistemática.

Seja v um modelo com p parâmetros, denominado por modelo reduzido, onde $v \subset \Omega$.

A deviance, também chamada razão de log-verossimilhanças, para comparar os referidos modelos, cuja distribuição pertence à família exponencial é

$$D = 2\log\left(\frac{L_\Omega}{L_v}\right) = 2\left(\log L_\Omega - \log L_v\right), \text{ em que} \qquad (27)$$

L_Ω indica o valor máximo da função de verossimilhança do modelo completo;

L_v indica o valor máximo da função de verossimilhança do modelo reduzido.

É importante acrescentar que comumente se multiplica a razão de log-verossimilhanças por 2, visto que $2\log\left(\frac{L_\Omega}{L_v}\right)$ resulta numa distribuição qui-quadrado, implicação interessante deste que é necessária uma distribuição amostral para determinar a região crítica da deviance (Dobson, 2002).

Assumindo que

$\tilde{\pi}_i = \tilde{\mu}_\Omega = y_i$ correspondem às estimativas do modelo completo;

$\hat{\pi}_i = \hat{\mu}_i$ representam as estimativas dos valores ajustados considerando o modelo v;

$a_i(\phi) = \phi/\lambda_i$, sendo ϕ um parâmetro de dispersão[3] constante e λ_i um peso a priori conhecido. Então, obtém-se a igualdade expressa por

$$2\log\left(\frac{L_\Omega}{L_v}\right) = 2\sum_{i=1}^{m}\frac{y_i(\tilde{\pi}_i - \hat{\pi}_i) - b(\tilde{\pi}_i) + b(\hat{\pi}_i)}{a_i(\phi)/\lambda_i}, \text{ tal que} \qquad (28)$$

$$D(y;\hat{\pi}_i) = 2\sum_{i=1}^{M}\lambda_i\left[y_i(\tilde{\pi}_i - \hat{\pi}_i) - b(\tilde{\pi}_i) + b(\hat{\pi}_i)\right], \qquad (29)$$

é a deviance do modelo reduzido em função apenas dos dados e das estimativas de máxima verossimilhança.

Com base nessas especificações, na modelagem de m ensaios de Bernoulli independentes, onde π_i indica a probabilidade de sucesso, é dito que,

$$\tilde{\pi}_i = \frac{y_i}{m_i},\ i = 1, 2, ..., M. \qquad (30)$$

A função de verossimilhança do modelo binomial é dada por

$$\prod_{i=1}^{M}\binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i - y_i}. \qquad (31)$$

---

[3] No contexto de modelos lineares generalizados tem-se $V(Y_i) = \phi V(\mu_i)$, ou seja, a variância é caracterizada por dois componentes: ϕ corresponde ao parâmetro de dispersão, independente da média, e $V(\mu_i)$ é a variância em função da média. Em particular, para o modelo binomial tem-se $\phi = 1$. No caso de $\phi > 1$, é dito existir efeito de superdispersão.

Ao ajustar um modelo logístico linear com parâmetros desconhecidos $\beta_1, \beta_2, ..., \beta_p$, os valores ajustados $\hat{\pi}_i$ são obtidos por

$$\text{logit}\left(\hat{\pi}_i\right) = \hat{\beta}_1 + \hat{\beta}_2 x_{i1} + ... + \hat{\beta}_p x_{ip} . \tag{32}$$

O logaritmo da função de verossimilhança maximizado para o modelo reduzido é calculado por

$$\log\left(\hat{L}_v\right) = \sum_{i=1}^{p}\left[y_i \log \hat{\pi}_i + \left(m_i - y_i\right)\log\left(1-\hat{\pi}_i\right) + \log\binom{m_i}{y_i}\right]. \tag{33}$$

No caso do modelo completo, as proporções ajustadas são as mesmas que as proporções observadas $\tilde{\pi}_i = \frac{y_i}{m_i}$.

Assim, dado que

$$\log\left(\hat{L}_\Omega\right) = \sum_{i=1}^{M}\left[y_i \log \tilde{\pi}_i + \left(m_i - y_i\right)\log\left(1-\tilde{\pi}_i\right) + \log\binom{m_i}{y_i}\right], \tag{34}$$

a deviance fica calculada por

$$D(y_i;\hat{\pi}_i) = 2\left[\log\left(\hat{L}_\Omega\right) - \log\left(\hat{L}_v\right)\right]$$
$$= 2\sum_{i=1}^{M}\left[y_i \log\left(\frac{\tilde{\pi}_i}{\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - \tilde{\pi}_i}{m_i - \hat{\pi}_i}\right)\right]. \qquad (35)$$

Segundo Lindsey (1997), a variável aleatória $D(y_i;\hat{\pi}_i)$ é distribuída aproximadamente como uma qui-quadrado χ^2_{m-p} com $m\text{-}p$ graus de liberdade, onde p é o número de parâmetros do modelo reduzido.

Quanto melhor for o ajuste do modelo linear generalizado aos dados, tanto menor será o valor da deviance D, isto, levando-se em consideração a obtenção de um modelo simples, que explique satisfatoriamente os dados, porém com desvios moderados, situados entre os modelos mais complicados e os que se ajustam mal aos dados (Cordeiro & Demétrio, 2007).

Caso isto não aconteça, existem fontes extras de variabilidade não contempladas pelo modelo. Algumas circunstâncias que podem contribuir para a falta de ajuste foram descritas por Hinde & Demétrio (1998), ou seja:

- alguma variável explanatória ou combinação delas foi omitida no preditor linear η;
- escolha inadequada da função de ligação $g(.)$;
- presença de *outliers* nos dados;
- fonte natural de variabilidade que a componente aleatória não consegue incorporar ao modelo, gerando o efeito da superdispersão.

A identificação e até mesmo o tratamento a ser dado para um *outlier* deve ser devidamente analisado. Com a finalidade de impedir que fontes extras de

variabilidade interfiram na qualidade do ajuste de um modelo (Hinde & Demétrio, 1998).

2.9 Resíduos

Em particular, os resíduos são usados para verificação de possíveis afastamentos das suposições feitas para o modelo. Os resíduos justificam a adequação ao modelo, além de proporcionar subsídio para que outras estatísticas sejam formuladas. Também, os resíduos são úteis para indicar a presença de observações discrepantes, que podem provocar interferência desproporcional nos resultados inferenciais (Turkman & Silva, 2000).

No contexto dos modelos lineares generalizados existem diversos tipos de resíduos propostos na literatura, dentre esses, os resíduos ordinários e a deviance residual.

Para todo o modelo ajustado, os resíduos na sua forma mais simples, conhecido como resíduo ordinário $\left(r_i\right)$ é dado por

$$r_i = y_i - \hat{\pi}_i \,. \qquad (36)$$

Na abordagem clássica de modelos de regressão binária, a detecção de pontos discrepantes é usualmente baseada no resíduo ordinário (36), sendo $\hat{\pi}_i$ a i-ésima proporção ajustada considerando o estimador de máxima verossimilhança.

A deviance D correspondente à diferença dos logaritmos das funções de verossimilhança observada e ajustada faz surgir uma nova definição de resíduo, conhecida como deviance residual. Pregibon (1981) define esse resíduo como

$$r_D = \text{sinal}\left(y_i - \hat{\pi}_i\right) d_i^{½}, \qquad (37)$$

sendo d_i a componente do desvio que mede a diferença dos logaritmos das funções de verossimilhança dos valores ajustados e observados, para a observação i-ésima correspondente.

Segundo Pregibon (1981), se existe uma transformação que normaliza a distribuição do resíduo (37) , então as raízes quadradas das componentes do desvio são resíduos que exibem as mesmas propriedades induzidas por esta transformação. Então, os resíduos r_D podem ser tratados como variáveis aleatórias tendo aproximadamente distribuição normal e, conseqüentemente, $r_D^2 = d_i$ tem aproximadamente distribuição χ_1^2.

Para o modelo binomial, segue a deviance residual correspondente.

$$r_D = \text{sinal}\left(y_i - \hat{\pi}_i\right)\left\{2\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\pi}_i}\right)\right]\right\}^{½}. \qquad (38)$$

Os resíduos da forma (37) apresentam vantagens ao não requerer o conhecimento da função normalizadora, exigirem computação mais simples após o ajuste do modelo linear generalizado (Cordeiro & Demétrio, 2007).

3 MATERIAL E MÉTODOS

3.1 Dados experimentais

Em consonância com o objetivo proposto neste trabalho, a aplicação da estratégia de alocação não proporcional para identificar *outliers* em dados binomiais foi realizada em um conjunto de dados (Tabela 3) provenientes de um experimento desenvolvido no Laboratório de Biologia do Departamento de Entomologia da ESALQ/USP.

Nesse experimento, teve-se o interesse de estudar o processo de competição de parasitas fêmeas *T. galloi* parasitando ovos de *A. kuehniella*, hospedeiro alternativo a *D. saccharalis*, praga de cana-de-açúcar, com o objetivo de obter o número de parasitas fêmeas que fornece a maior proporção de ovos parasitados. Maiores detalhes podem ser encontrados em Borgatto (2004).

TABELA 3 Dados referente ao experimento para medir a competição de fêmeas *T. galloi* sob um número fixo de 128 ovos de *A. kuehniella.*

ovos parasitados	número total de ovos	número de fêmeas
0	128	8
0	128	8
19	128	8
49	128	8
22	128	8

... continua ...

"TABELA 3, Cont."

51	128	8
65	128	8
0	128	16
0	128	16
57	128	16
35	128	16
52	128	16
37	128	16
58	128	16
0	128	32
0	128	32
120	128	32
90	128	32
86	128	32
102	128	32
95	128	32
0	128	64
0	128	64
0	128	64
57	128	64
77	128	64
105	128	64
99	128	64
0	128	128
38	128	128
21	128	128
82	128	128
42	128	128
90	128	128
81	128	128

Convém salientar que para fins didáticos, algumas respostas foram excluídas com a finalidade de impor uma alta variabilidade resultante da presença de observações que diferem em magnitude das demais observações em cada grupo (número de fêmeas) e que poderão ser consideradas *outliers* no conjunto de dados.

Esse fato foi confirmado pelo gráfico boxplot apresentado na seção de resultados e discussão.

3.2 Modelos avaliados

Ajustou-se um modelo binomial com uso das funções de ligação logit, probit e complemento log-log, dadas nas equações (9), (10) e (11). Nos modelos ajustados, c representou o intercepto, π_{ij} referiu-se ao valor da proporção ajustada, η_{ij} foi o preditor linear em que s_i correspondeu ao parâmetro do i-ésimo grupo (número de fêmeas). Assim, para cada função de ligação, assumindo a observação (y_{ij}) como o número de casos de ovos parasitados na j-ésima repetição (j = 1,..., 7) do i-ésimo grupo de fêmeas (i=1,...,5), têm-se os seguintes modelos.

- **logit**

$$E\left(\frac{y_{ij}}{m_i}\right) = \frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}} = \pi_{ij}, \text{ em que} \qquad (39)$$

$$\eta_{ij} = \ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = c + s_i ;$$

- **probit**

$$E\left(\frac{y_{ij}}{m_i}\right) = \Phi\left(\eta_{ij}\right) = \pi_{ij}\text{, em que} \tag{40}$$

$$\eta_{ij} = \Phi^{-1}\left(\pi_{ij}\right) = c + s_i;$$

- **complemento log-log**

$$E\left(\frac{y_{ij}}{m_i}\right) = 1 - e^{e^{\eta_{ij}}} = \pi_{ij}\text{, em que} \tag{41}$$

$$\eta_{ij} = \log\left[-\log\left(1-\pi_{ij}\right)\right] = c + s_i.$$

Em todos os modelos m_i foi fixado em 128, conforme pode ser verificado na Tabela 3.

3.3 Adaptação do método de alocação não-proporcional para dados binários

Para efetuar os passos necessários à execução do método proposto para identificar *outliers* em dados binomiais, utilizou-se de recursos de programação disponível no R (R Development Core Team, 2007). O programa implementado (Anexo A) utilizou procedimentos da função glm (generalized linear model).

A descrição do algoritmo utilizado para a alocação não-proporcional procede de acordo com os seguintes passos:

Etapa 1: Escolha do conjunto inicial **S.**

Determinou-se um subconjunto das unidades amostrais. Para isso, foi selecionada de cada grupo de fêmeas a observação que respeitou a seguinte condição

$$\min \left| y_{ij} - \text{med}_i \right|, \; j = 1,...7 \text{ e } i = 1,...,5. \tag{42}$$

Após a seleção do primeiro subconjunto obtido conforme condição apresentada em (42), a identificação dos *outliers* teve como ponto de partida o ajuste dos modelos a que se refere à seção 3.2, para o primeiro subgrupo composto por algumas observações selecionadas. A continuidade da seleção de outras observações foi conduzida pela estratégia denominada como alocação não-proporcional.

Etapa 2: Adicionando observações durante a busca

Após o ajuste dos modelos (39), (40) e (41), considerando o conjunto inicial **S**, ordenou-se de modo decrescente as observações dentro de cada grupo (número de fêmeas), tendo como referência o quadrado da diferença das proporções observadas e ajustadas definidas por

$$\text{dif} = \left[\frac{y_{ij}}{m_i} - \hat{\pi}_{ij} \right]^2 . \tag{43}$$

Em seguida, selecionou-se a observação que apresentou o menor quadrado na expressão (43). Na realização dessa etapa, originou-se um novo conjunto de observações, denominado por $S^{(a+1)}$. Assumindo este novo conjunto,

repetiu-se novamente esse processo. O procedimento finalizou quando todas as observações foram inseridas em cada grupo (número de fêmeas).

Etapa 3: Estimativas obtidas na inserção de *outliers*

Para cada subconjunto obtido em $S^{(a+1)}$, computaram-se as estimativas de máxima verossimilhança dos parâmetros do modelo binomial, com as funções de ligação citadas na seção 3.1.

Dando seqüência ao estudo, a cada novo subconjunto obtido, foi computado o valor das estatísticas de interesse geradas pelo modelo ajustado.

Por fim, a interpretação dos recursos foi realizada por meio de uma inspeção nos gráficos, conforme é ilustrado na seção de resultados e discussão.

4 RESULTADOS E DISCUSSÃO

Uma análise preliminar das observações referente ao número de ovos parasitados em cada grupo (número de fêmeas) do experimento foi realizada conforme ilustra o gráfico boxplot (Figura 3).

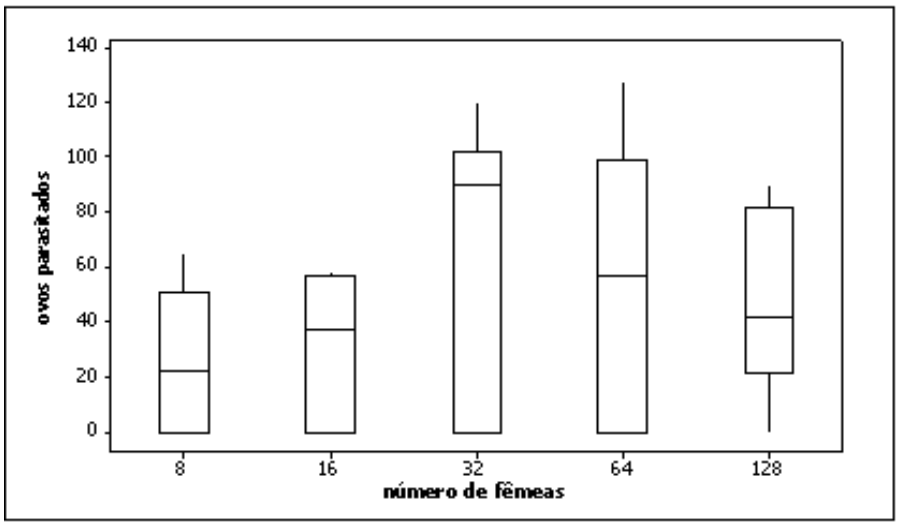


FIGURA 3 Boxplot dos números de ovos parasitados de *A. kuchniella* por fêmeas *T. galloi* no experimento realizado no Laboratório de Biologia do Departamento de Entomologia da ESALQ/USP .

Analisando os resultados ilustrados pela Figura 3, pode-se observar que os grupos (número de fêmeas 32 e 64) indicam a existência de alta variabilidade dos dados quando comparado com os demais. Assim, os dados experimentais, possivelmente, apresentam evidências da falta de ajuste ao considerar um modelo linear generalizado. Dessa forma, a adoção de estratégias que

possibilitam alocações de unidades amostrais, torna-se um instrumento interessante para este estudo.

Neste contexto, o presente trabalho propõe estender o método da alocação não-proporcional (Bertaccini & Varriale, 2006), considerando os dados (Tabela 3) descritos na seção 3.1, para o modelo binomial com diferentes funções de ligação. Em síntese, a execução desse método se justifica porque na alocação não-proporcional, a cada passo, uma unidade amostral é selecionada em conformidade com a condição (43), evitando a entrada de *outliers* nos subconjuntos amostrais dos passos iniciais, propiciando um monitoramento da dispersão dos dados para cada estatística de interesse calculada.

Mantendo o enfoque nos subconjuntos formados a partir dos passos 16º a 22º (Tabela 4), realizou-se a análise do efeito dos *outliers* em relação às estimativas dos parâmetros dos modelos propostos (Seção 3.2), usando as funções de ligação logit, probit e complemento log-log. Os resultados encontrados nas Figuras 4a, 4b e 4c mostraram que ao decorrer dos primeiros 16 passos, as estimativas dos coeficientes apresentaram comportamento semelhante, independente da função de ligação utilizada, indicando a ausência de *outliers* nos respectivos subconjuntos dos dados amostrais.

Alguns subgrupos obtidos da aplicação deste método, encontram-se descritos na Tabela 4, sendo que o número de elementos considerados em cada subconjunto amostral S, identificados pelos passos 16º ao 22º, apresentou observações que revelaram ser supostamente *outliers*, visto que em algumas repetições, as observações diferiram em magnitude, comparando-se com as demais.

TABELA 4 Subgrupos das observações referentes ao número de ovos parasitados nos subgrupos de fêmeas, formados nos 16º a 22º passos usando método de inspeção para identificar *outliers* em dados binomiais por meio da estratégia não-proporcional.

grupos (nº de fêmeas)	**16º passo**	**grupos (nº de fêmeas)**	**22° passo**
8	22, 19	**8**	0, 0, 22, 19
16	35, 37, 52, 57, 58	**16**	35, 37, 52, 57, 58
32	86, 90, 95,102	**32**	86, 90, 95,102
64	57, 77	**64**	57, 77
128	21, 38, 42	**128**	0, 21, 38, 42

Entretanto, nos resultados encontrados na Tabela 4, os subconjuntos gerados nos passos 16º a 22º obtiveram observações com valores muito distantes das outras já inseridas, e consequentemente expressaram estimativas diferenciadas, conforme mostrado pelas Figuras 4a, 4b e 4c, revelando que foram afetadas pela entrada de *outliers* na formação dos subgrupos.

Convém salientar, que nos subconjuntos formados no passo 22º, nota-se pelo número de observações que além de existir observações discrepantes, a sub-amostra evidencia um inflacionamento de zeros. Com isto, pode-se verificar que a extensão do método para o modelo binomial com aplicação da estratégia não-proporcional pode ser recomendável e viável para identificar *outliers*.

Comparando-se as diferentes funções de ligação, verifica-se que os modelos ajustados em todos os grupos foram equivalentes na identificação dos

outliers, ou seja, para este estudo a escolha da função de ligação não foi um fator que propiciou vantagens na identificação de *outliers*.

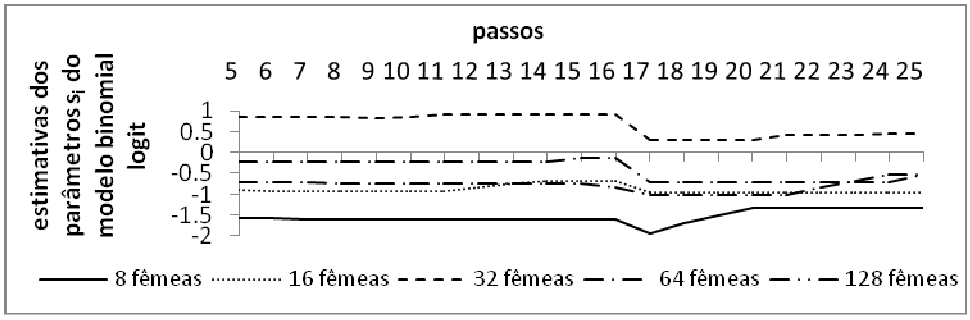
Entretanto, em termos de resultados numéricos, ressalta-se que o uso das funções logit, probit e complemento log-log apresentaram estimativas para os parâmetros do modelo binomial muito próximas, conforme pode ser constatado pela semelhança dos gráficos (Figuras 4a, 4b e 4c).

A partir dos passos próximos ao 22º, a inserção dos demais *outliers* possibilitou que o conjunto de dados adquirisse certa uniformidade, de tal modo que as estimativas dos parâmetros do modelo retomaram valores com pequenas oscilações, mascarando a existência de observações atípicas no conjunto de dados.

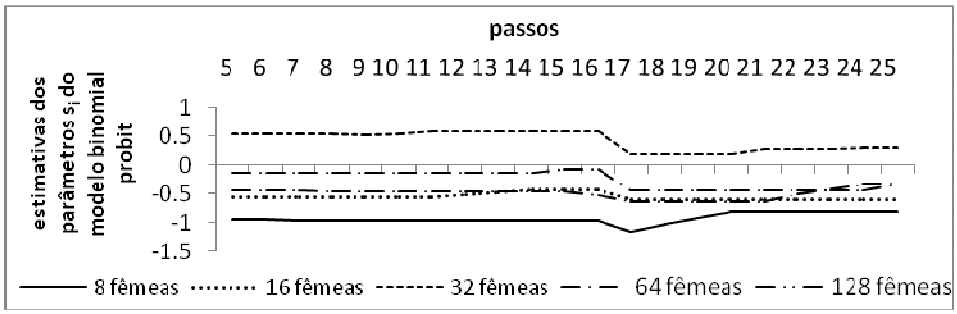
Assim, o comportamento revelado pelos gráficos nos passos finais sugeriu a falsa impressão de ausência da notável variabilidade, apresentando estimativas influenciadas pelos demais *outliers* inseridos nos subconjuntos das observações dentro de cada grupo.

Analogamente, o mesmo comportamento foi observado nas situações em que a estimativa do parâmetro do primeiro grupo (8 fêmeas) foi confundida com as estimativas dos demais efeitos. Os resultados podem ser vistos na Figura 5 e confrontados com os resultados ilustrados na Figura 4.

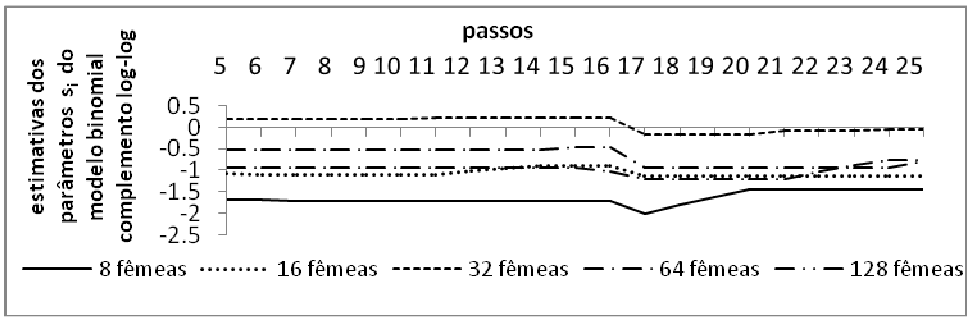
A partir dos resultados da Figura 6, verificou-se que a deviance apresentou valores que não diferiram entre os modelos. É mostrado que existem fortes evidências para que o modelo seja rejeitado, visto que a deviance foi determinada por altos valores em relação ao número de graus de liberdade, sinalizando a existência de variabilidade acentuada nos subgrupos amostrais.



(a)

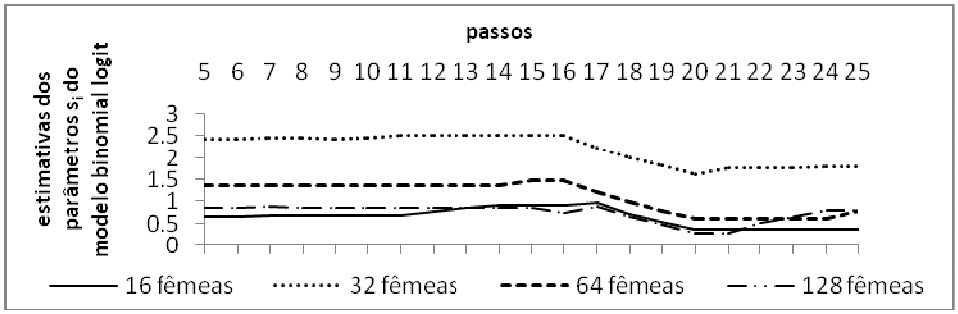


(b)

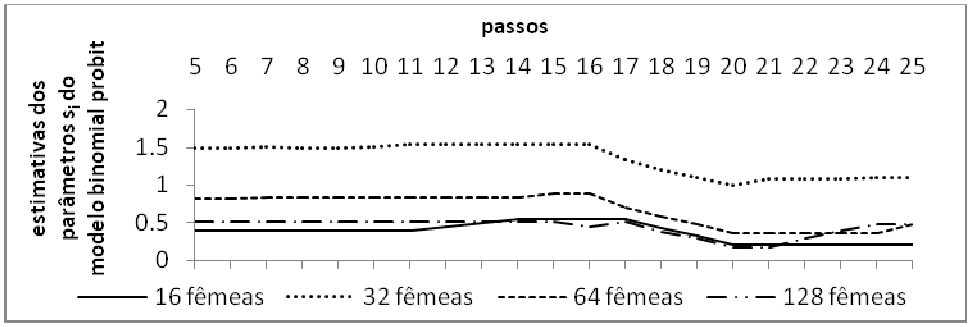


(c)

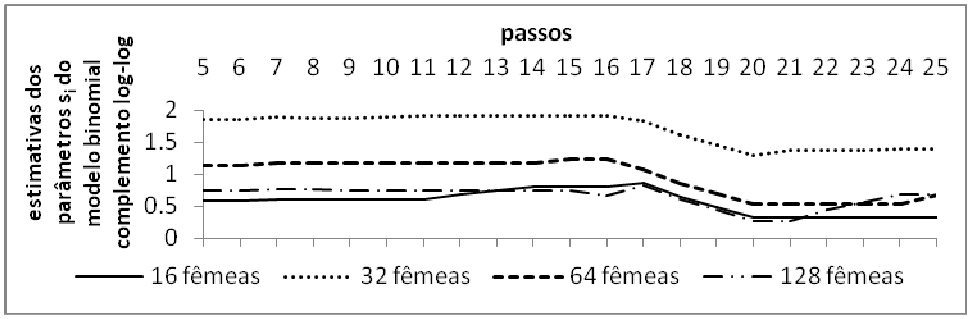
FIGURA 4 Inspeção das estimativas dos parâmetros s_i do modelo binomial sem o intercepto c, com funções de ligação logit (a), probit (b) e complemento log-log (c), utilizado para modelar a proporção de ovos parasitados a cada passo efetuado do método de inspeção com a estratégia não-proporcional.



(a)



(b)



(c)

FIGURA 5 Inspeção das estimativas dos parâmetros do modelo binomial, considerando o intercepto c e as funções de ligação logit (a), probit (b) e complemento loglog (c), na modelagem da proporção de ovos parasitados a cada passo efetuado do método com a estratégia não-proporcional.

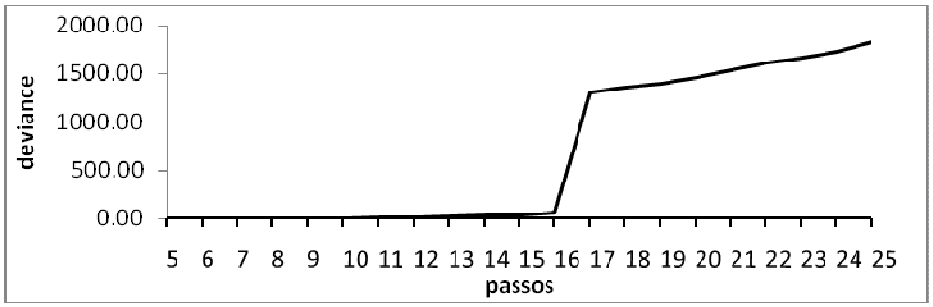


FIGURA 6 Inspeção dos valores da deviance do modelo binomial considerando as funções de ligação logit, probit complemento log-log para modelar a proporção de ovos parasitados a cada passo efetuado do método de inspeção com a estratégia não-proporcional.

Os resultados ilustrados na Figura 7 apresentaram o comportamento das três funções de ligação usadas neste trabalho para os modelos ajustados com presença ou ausência do intercepto c. Observou-se que as proporções ajustadas $\hat{\pi}$, cujos valores são próximos a zero, tem preditor linear superior quando se trata da função probit, enquanto que a logit e a complemento log-log apresentam comportamento semelhante, resultado já previsto pela literatura, conforme comentado na seção 2.6. À medida que os valores ajustados aproximam-se de 1, a função logit tem crescimento mais acentuado, principalmente quando comparada com a complemento log-log (McCullagh & Nelder, 1989).

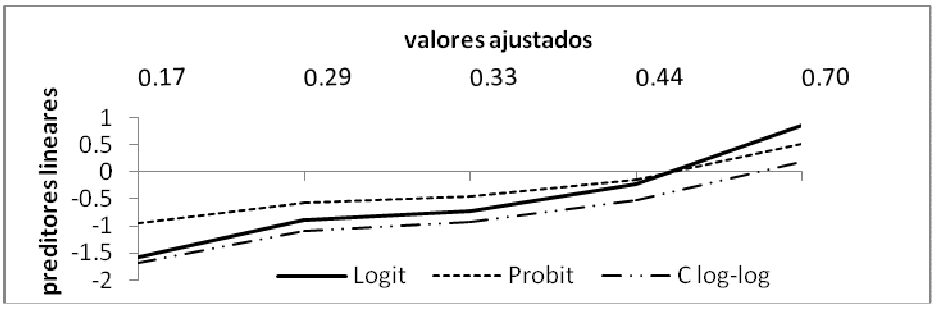


FIGURA 7 Valores ajustados das proporções do modelo binomial usando funções de ligação logit, probit e complemento log-log utilizado para modelar a proporção de ovos parasitados, a cada passo efetuado do método de inspeção com a estratégia não-proporcional.

5 CONCLUSÕES

A metodologia proposta neste estudo para identificação de *outliers* com aplicação da estratégia de alocação não-proporcional foi viável e recomendável para dados binomiais, sendo portanto, uma importante técnica para análise dos dados quando se tem interesse em verificar a presença de *outliers*.

A aplicação da estratégia de alocação não-proporcional para identificar *outliers* em dados binomiais detecta a presença de observações atípicas de modo semelhante para as funções de ligação logit, probit ou complemento log-log. Então, caberá ao pesquisador verificar qual função se enquadra de modo mais satisfatório à natureza do conjunto de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, A.C.G. **Efeitos de especificação incorreta da função de ligação no modelo de regressão beta**. 2007. 88p. Dissertação (Mestrado em Estatística) – Universidade de São Paulo, São Paulo.

ATKINSON, A.C.; RIANI, M. **Robust diagnostic regression analysis**. New York: Springer, 2000.

BARNETT, V. Outlier and order statistics. **Communications in Statistics**: part A: theory and methods, New York, v. 17, n. 7, p. 2109–2118, 1988.

BARNETT, V.; LEWIS, T. **Outliers in statistics**. 3.ed. New York: J. Wiley, 1993.

BERTACCINI, B.; VARRIALE, R. Robust Analysis of variance: an approach based on the Forward Search. **Computational Statistics & Data Analysis**, Amsterdam, v.51, n.10, p.5172-5183, June 2006.

BORGATTO, A.F. **Modelos para proporções com superdispersão e excesso de zeros**: um procedimento bayesiano. 2004. 90p. Tese (Doutorado em Estatística com Experimentação Agronômica)–Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, São Paulo.

CORDEIRO, G.M.; DEMÉTRIO, C.G.B. Modelos lineares generalizados. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 12., 2007, Santa Maria. **Anais...** Santa Maria: UFSM, 2007. 1 CD-ROM.

CORDEIRO, G.M.; LIMA NETO, E.A. Modelos paramétricos. In: SIMPÓSIO BRASILEIRO DE PROBABILIDADE E ESTATÍSTICA, 16., 2006, Caxambu. **Anais...** São Paulo: Associação Brasileira de Estatística, 2006. 1 CD-ROM.

DOBSON, A. J. **An introduction to generalized linear models**. 2. ed. London: Chapman & Hall, 2002. 225p.

HINDE, J.; DEMÉTRIO, C.G.B. Overdispersion: models and estimation. **Computational Statistics & Data Analysis**, Amsterdam, v.27, n.2, p.151-170, Apr.1998.

LINDSEY, J.K. **Applying generalized linear models**. New York: Springer, 1997. 256p.

MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. **Introduction to the theory of statistics**. 3.ed. New York: McGraw-Hill, 1974. 564p.

MYERS, R.H.; MONTGOMERY, D.C.; VINING, G.G. **Generalized linear models:** with applications in engineering and the sciences. New York: J. Wiley, 2002. 346p.

PREGIBON, D. Logist regression diagnostics. **Annals of Statistics**, Hayward, v.9, n.4, p. 705-724, July 1981.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2007.

TUKEY, J.W. A survey of sampling from contaminated distribution. **Contributions to probability and statistics.** California: University Stanford, 1960.

TURKMAN, M.A.A.; SILVA, G.L. **Modelos lineares generalizados**: da teoria à prática. Lisboa: Sociedade Portuguesa de Estatística, 2000. 151p.

ANEXO

Página

PROGRAMA: Programa implementado no R (R Development Core Team, 2007) para seleção das observações, passos a passo, e obtenção das estimativas dos parâmetros de interesse do modelo binomial.

```
# ### Leitura e montagem do banco de dados ###### #

dados=read.table("dados.txt",h=T)
propobs=dados$obs/dados$total
ind=c(rep(0,nrow(dados)))
mdados=cbind(dados$ano,dados$est,dados$total,dados$obs,propobs,ind)

# ##### Variáveis a serem declaradas ############ #

tsub=27
nsub=4

difmed=matrix(0,tsub,1)
forma_conj=matrix(0,108,108)
conj=matrix(0,1,7)
m2=matrix(0,1,7)
selajus=matrix(0,1,2)

######################### Funções ################## #

### Função cálculo das diferenças de médias ######

dmed=function(matmed)
{

# ### Calculo da diferença dos grupos ############

conta1=1
conta2=tsub
  for (c1 in 1:nsub)
{
   med=aux[c1,1]
   auxmed=rep(med,tsub)
   dife=abs(mdados[conta1:conta2,4]-auxmed)
```

```
   tdife=as.matrix(dife)
   conta1=conta1+tsub
   conta2=conta2+tsub

   difmed=rbind(difmed,tdife)
 }
   difmed=difmed[28:135,1]
   mdados=cbind(mdados,difmed)

return(mdados)
}

# ######### seleciona valores ####################

seleciona=function(dad,nsub)
{

conta1=1
conta2=tsub

for (c2 in 1:nsub)

{
  num=min(dad[conta1:conta2,7])

 for (c3 in 1:nrow(dad))

{
   if (dad[c3,7]==num)

      {
        dad[c3,6]=1
        daux=dad[c3,]
        conj=rbind(conj,daux)
      }
}
      conta1=conta1+tsub
      conta2=conta2+tsub

}
```

```
   conj=conj[2:nrow(conj),]
   return(list(dadsel=conj,dadmarc=dad))
}

ajus=function(auxconj1)
{
   y=auxconj1[,4]
   f=auxconj1[,3]-auxconj1[,4]
   ano=factor(auxconj1[,1])
   resp=cbind(y,f)
   modelo=glm(resp~-1+ano,family=binomial)
   combpred=cbind(auxconj1[,1],modelo$fitted.values)

   estcof=modelo$fitted.values
   devres=modelo$linear.predictors
   devnul=modelo$null.deviance

    return(list(resajus=combpred,cof=estcof,devr=devres,devn=devnul))

}
residuo=function(aux2conj1,obsajus,nsub,tsub)
{
  for (c1 in 1:nrow(aux2conj1))
 {
   if (aux2conj1[c1,6]==0)
     {
      maux=t(as.matrix(aux2conj1[c1,]))
      m2=rbind(m2,maux)
     }
  }
      m2=m2[2:nrow(m2),]
      cap2=0
      resi=matrix(0,nrow(m2),1)

  for (c2 in 1:nrow(obsajus))
  {
     cap1=obsajus[c2,1]

   if (cap1!=cap2)
```

```
    {
     cap2=obsajus[c2,1]
     aj=obsajus[c2,]
     selajus=rbind(selajus,aj)
    }
    }

selajus=selajus[2:nrow(selajus),]
for (c3 in 1:nrow(selajus))

 {
   x=t(as.matrix(selajus[c3,]))

for (c4 in 1:nrow(m2))

 {

 if (m2[c4,1]==x[1,1])

 {
   resi[c4,1]=(x[1,2]-m2[c4,5])^2

 }

 }

 }

 m2=cbind(m2,resi)
 minimo=as.real(min(resi[,1]))

 for (i5 in 1:nrow(m2))

 {
    if (m2[i5,8]==minimo) obsmim=m2[i5,]
 }

return(list(e1=m2,e2=selajus,e3=resi,e4=minimo,e5=obsmim))

}
```

```
atual=function(aux2conj1,auxconj1,obs)

 {
   for (i6 in 1:nrow(aux2conj1))
  {
    if (aux2conj1[i6,5]==obs[1,5])

       {
         aux2conj1[i6,6]=1
         auxconj1=rbind(auxconj1,aux2conj1[i6,])
       }

  }

   atuamarc=auxconj1
   dadatual=aux2conj1

  return(list(a1=dadatual,a2=atuamarc))
}

# ############ PROGRAMA PRINCIPAL ##################


# ####### Construção do primeiro sub-conjunto #############

  med<-tapply(dados$obs,list(dados$ano),median)
  aux=as.matrix(med)

  tabela=dmed(aux)
  conj1=seleciona(tabela,nsub)

  ds=conj1$dadsel
  dm=conj1$dadmarc
  contlin=nrow(dm)
  obspred=ajus(ds)
  auxobspred=obspred$resajus


  # ##### coleta de estatísticas #######
```

```
 est_coef=t(as.matrix(obspred$cof))
 est_devr=t(as.matrix(obspred$devr))
 est_devh=t(as.matrix(obspred$devn))

saida=residuo(dm,auxobspred,nsub,tsub)

# ###### Fim do primeiro conjunto ######################

parada=1
while (parada<=contlin)

  {
  jobs=t(as.matrix(saida$e5))
  jobs=t(as.matrix(jobs[,1:7]))

  chatual=atual(dm,ds,jobs)
  parada=parada+1

  ds=chatual$a2
  dm=chatual$a1
  obspred=ajus(ds)
  auxobspred=obspred$resajus

 est_coef=rbind(est_coef,obspred$cof)
 est_devr=rbind(est_devr,obspred$devr)
 est_devh=rbind(est_devh,obspred$devn)

saida=residuo(dm,auxobspred,nsub,tsub)

}
 est_coef
 est_devr
 est_devh
```