GEMMA LUCIA DUBOC DE ARAUJO

MÉTODOS DE ESTIMAÇÃO EM REGRESSÃO LOGÍSTICA COM EFEITO ALEATÓRIO: APLICAÇÃO EM GERMINAÇÃO DE SEMENTES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2012

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
A663m
2012

Araujo, Gemma Lucia Duboc de, 1961-
Métodos de estimação em regressão logística com efeito aleatório: aplicação em germinação de sementes / Gemma Lucia Duboc de Araujo. – Viçosa, MG, 2012.
xii, 94f. : il. (algumas col.) ; 29cm.

Inclui apêndice.
Orientador: Sebastião Martins Filho.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Modelos lineares (Estatística). 2. Análise multivariada. 3. Análise de regressão. 4. Famílias exponenciais (Estatística). 5. Estimativa de parâmetro. 6. Pinhão-manso. I. Universidade Federal de Viçosa. II. Título.

CDD 22. ed. 519.5

GEMMA LUCIA DUBOC DE ARAUJO

MÉTODOS DE ESTIMAÇÃO EM REGRESSÃO LOGÍSTICA COM EFEITO ALEATÓRIO: APLICAÇÃO EM GERMINAÇÃO DE SEMENTES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 1º de fevereiro de 2012.

Jaques Silveira Lopes

Marcos Deon Vilela de Resende

Fabyano Fonseca e Silva
(Coorientador)

Sebastião Martins Filho
(Orientador)

A meus pais, Joseph e Amelia Iris.

A meus filhos, Priscila, Fabio, Kim Flavio e Sophia.

A meus irmãos, cunhados e sobrinhos.

À Irmã Elisa, à Irmã Hildete e à Irmã Mercês.

A meus novos, fiéis e torcedores amigos do mestrado.

Dedico.

AGRADECIMENTOS

Agradeço a Deus que guiou minha vida para que eu chegasse até aqui, que fez de todos os contratempos uma maneira de me fortalecer e de me dar amigos fiéis; que me fez perceber que Nele tudo posso, ainda que tudo diga o contrário.

Agradeço a Maria Santíssima, ao meu Anjo da Guarda e a todos os anjos e santos a quem pedimos que intercedessem ao Pai por mim.

Agradeço a meus pais, pela ajuda moral, financeira, prática e com as orações sem cessar.

Agradeço a todos os meus familiares e amigos que rezaram e torceram por mim.

Agradeço à Margareth e à Rosane, colegas e amigas do Dept^{o} de Matemática da UFV, pelo carinho, atenção e incentivo.

Agradeço a todos aqueles que me emprestaram seus ouvidos e seus ombros…

Agradeço a todos os meus mestres do Dept^{o} de Estatística da UFV, que direta ou indiretamente colaboraram para a minha formação intelectual.

Agradeço a todos os meus colegas de mestrado, que foram mais que colegas, foram companheiros e amigos.

Agradeço ao André, amigo e irmão de orientação, que com toda a paciência e entusiasmo me introduziu no mundo da programação estatística, prestando apoio incondicional.

Agradeço ao Prof. Fabyano por sua disponibilidade e atenção.

Agradeço especialmente ao Prof. Sebastião, que se faz presente em quase todos os itens desta lista.

Por fim, agradeço a todos aqueles que foram meus amigos invisíveis...

A todos, muito obrigado!

"O último passo da razão está em admitir que há infinitas coisas que ultrapassam o seu alcance; se a isto não chegar, dá prova de grande fraqueza." (Pascal)

BIOGRAFIA

GEMMA LUCIA DUBOC DE ARAUJO, filha de Amelia Iris Duboc de Araujo e Joseph Ildefonso de Araujo, nasceu em 12 de março de 1961, em Valença-RJ.

Em dezembro de 1984 graduou-se em Licenciatura em Ciências e Licenciatura Plena em Matemática pela Universidade Federal de Viçosa.

Em 1996, concluiu pós-graduação *lato-sensu* em "Construção do conhecimento e Currículo" pela Universidade Federal Fluminense.

Em 2002, concluiu pós-graduação *lato-sensu* em "Ciência da Computação" pela Universidade Federal de Viçosa.

Em março de 2009, iniciou no Programa de Pós-graduação, a nível de Mestrado em Estatística Aplicada e Biometria, na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 1º de fevereiro de 2012.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BLUE	*best linear umbiased estimator* (melhor estimador linear não-viesado)
BLUP	*best linear umbiased predictor* (melhor preditor linear não-viesado)
EMQ	estimativa de quadrados mínimos
EMV	estimativa de máxima verossimilhança
EQM	erro quadrático médio
GLM	modelos lineares generalizados
GLMM	modelos lineares generalizados mistos
IOR	intervalo percentílico da razão de chances mediana
IWLS	*iterated weighted least squares* (método dos quadrados mínimos ponderados iterativo)
ML	*maximum likelihood* (máxima verossimilhança)
MNR	método de Newton-Raphson
MOR	*median odds ratio* (razão de chances mediana)
OR	*odds ratio* (razão de chances, risco)
REML	*restricted maximum likelihood* (máxima verossimilhança restrita)
TRV	teste da razão de verossimilhança

RESUMO

ARAUJO, Gemma Lucia Duboc de, M.Sc., Universidade Federal de Viçosa, fevereiro de 2012. **Métodos de estimação em regressão logística com efeito aleatório: aplicação em germinação de sementes.** Orientador: Sebastião Martins Filho. Coorientadores: Enrico Antônio Colosimo e Fabyano Fonseca e Silva.

Em modelos de regressão logística a inclusão do efeito aleatório no intercepto permite capturar os efeitos de fontes de variação provenientes das características particulares de um grupo (heterogeneidade), desinflacionando o erro puro e provocando uma flutuação no intercepto do modelo. Esta inclusão traz complexidade nos métodos de estimação e também muda a interpretação dos parâmetros que, dada originalmente pela razão de chances, passa a ser vista sob o enfoque da razão de chances mediana. A estimação dos parâmetros de um modelo misto pode ser feita por muitos métodos diferentes com desempenho variado, como o método da aproximação de Laplace, da máxima verossimilhança (ML) e da máxima verossimilhança restrita (REML). Assim, o objetivo deste trabalho foi verificar em modelos de regressão logística com efeito aleatório no intercepto as consequências na interpretação dos parâmetros, na qualidade de um experimento e na classificação de tratamentos via razão de chances mediana, e verificar o desempenho dos métodos de estimação acima citados. As análises foram feitas sob simulação e posteriormente num conjunto de dados reais de um experimento com germinação de sementes de pinhão-manso (*Jatropha curcas* L.). Considerando o modelo de regressão logística com efeito aleatório no intercepto, verificou-se que o método de estimação REML apresentou melhor desempenho e que a variância do efeito aleatório afeta o desempenho de qualquer um dos métodos avaliados sendo estes inversamente proporcionais. Sugerem-se novos estudos para determinar com mais propriedade a influência dos pontos de estabilização e do nível mediano de efetividade na eficiência dos métodos. No experimento de avaliação de germinação de sementes de pinhão-manso envolvendo os substratos rolo de papel, sobre papel, sobre areia e entre areia, a inclusão do efeito aleatório no modelo logístico apontou considerável heterogeneidade na germinação de sementes em unidades diferentes de um mesmo substrato. A razão de chances mediana apontou a superioridade do substrato entre areia em relação a sobre papel na germinação de sementes de pinhão-manso, resultado semelhante ao obtido pelo teste de Tukey.

ABSTRACT

ARAUJO, Gemma Lucia Duboc de, M.Sc. Universidade Federal de Viçosa, February, 2012. **Estimation methods in logistic regression with random effects: application in seed germination.** Adviser: Sebastião Martins Filho. Co-Advisers: Enrico Antônio Colosimo and Fabyano Fonseca e Silva.

In logistic mixed models with random effect on intercept allows capturing the effects of sources of variation from the particular characteristics of a group (heterogeneity), deflating the pure error and causing a fluctuation in the model intercept. This inclusion brings complexity in estimation methods and also changes the interpretation of the parameters that, originally given by the odds ratio, is then seen from the median odds ratio. The estimation parameters of a mixed model can be made by many different methods with varying performance, as the Laplace's approximation method, maximum likelihood (ML) and restricted maximum likelihood (REML). The objective of this work was to verify in logistic mixed models with random effects on intercept the consequences in interpretation of parameters, in quality of experiment and in classification of treatment via the median odds ratio, and verify the performance of the estimation methods above cited. The analyzes were performed under simulation and after in set of real data from seeds germination experiment of physic nut (*Jatropha curcas* L.). Considering the logistic mixed model with random effects on intercept, it was verified that the REML estimation method performed better and that the variance of the random effect affects the performance of any of these methods being evaluated inversely proportional. We suggest further studies to determine more properly the influence of the inflexion points and the effective median level in performance methods. In the experiment to evaluate the seeds germination of physic nut involving roll paper, on paper, on sand and between sand substrates, the inclusion of random effects in logistic model showed considerable heterogeneity in seeds germination in different units of the same substrate. The median odds ratio showed the superiority of the substrate between sand over on paper in seeds germination of physic nut, result similar to that obtained by the Tukey's test.

INTRODUÇÃO GERAL

Em estudos de diversas áreas como saúde, economia, psicologia, recursos humanos, agronomia etc., muitas vezes a variável resposta apresenta apenas duas categorias que podem ser expressas por sim ou não, presença ou ausência, sucesso ou fracasso, e outras variações destas respostas que, por sua vez, podem ser traduzidas por 1 ou 0. Essas variáveis são ditas dicotômicas ou binárias, e possuem distribuição de Bernoulli, de forma que a modelagem da relação entre estas e outras variáveis explicativas – que podem ser contínuas ou discretas – se faz por meio da regressão logística, a qual é uma classe dos modelos lineares generalizados.

Embora o modelo logístico seja não linear, é possível linearizá-lo mediante uma transformação na variável resposta, chamada logito (ou *logit*), a qual é denominada função de ligação. Essa linearização permite uma interpretação simples e objetiva dos parâmetros regressores por meio do conceito de razão de chances (*odds ratio*).

Um problema que pode ser considerado dentro do estudo de regressão logística é a incorporação de efeitos aleatórios ao modelo, os quais, segundo Larsen et al. (2000), refletem a heterogeneidade dentro dos níveis das variáveis explicativas (tratamentos). A inclusão dos efeitos aleatórios em experimentos com dados correlacionados é importante porque permite recuperar a informação entre tratamentos além de desinflacionar o erro puro (WEBER, 2011). Porém, a inclusão de tais efeitos proporciona um aumento significativo na complexidade do modelo, havendo a necessidade de optar por métodos de estimação adequados a tal condição. Além disso, a inclusão do efeito aleatório dificulta a interpretação dos resultados, uma vez que interfere na interpretação dos parâmetros de efeito fixo.

Diante do exposto, objetivou-se estudar o modelo de regressão logística com efeito aleatório introduzido no intercepto com ênfase na comparação da eficiência de diferentes métodos de estimação (máxima verossimilhança, máxima verossimilhança restrita e aproximação de Laplace) e na interpretação da razão de chances. Para tanto, utilizaram-se dados simulados e dados reais provenientes de um experimento para avaliação de germinação de sementes de pinhão-manso (*Jatropha curcas* L.).

O presente trabalho está organizado da seguinte maneira: no capítulo 1 é apresentada a teoria básica dos modelos lineares generalizados; no capítulo 2 são apresentados os conceitos, a estimação e a interpretação dos parâmetros do modelo de regressão logística; no capítulo 3 é apresentado o modelo logístico com efeito aleatório

e suas consequências na interpretação dos parâmetros; no capítulo 4 encontra-se um artigo contendo alguns métodos de estimação utilizados para o modelo apresentado no capítulo 3 bem como estratégias de comparação de tais métodos e estudos de simulação; finalmente, no capítulo 5 encontra-se um artigo no qual é realizada uma aplicação de toda metodologia abordada nos capítulos anteriores a um conjunto de dados reais provenientes de um experimento de germinação de sementes de pinhão-manso (*Jatropha curcas* L.), com ênfase na razão de chances mediana.

CAPÍTULO 1 – MODELOS LINEARES GENERALIZADOS

Muitos métodos de estimação paramétrica assumem a normalidade dos dados. No entanto, nem sempre esta suposição é plausível e a utilização de tais métodos leva a um desempenho insatisfatório (SSI, 2011). Assim, a busca de novos métodos tornou-se imperativa e até o início do século XX os métodos desenvolvidos para respostas contínuas atendendo a outros tipos de distribuição já haviam ganhado um bom nível de sofisticação. O que não aconteceu com os métodos para respostas categóricas, apesar do importante trabalho de Pearson, por volta de 1900. A partir de então e até a década de 60 foram desenvolvidos alguns modelos e técnicas de regressão para variáveis categóricas atendendo a vários tipos de dados (AGRESTI, 2002). Foi, porém, na década de 70 que o avanço computacional permitiu o uso de métodos iterativos, levando a novos progressos na estimação de parâmetros em modelos de regressão para diversas distribuições de variáveis (PAULA, 2004).

Em 1972, Nelder e Wedderburn apresentaram uma proposta inovadora, unificando as teorias de modelagem estatística através de uma classe de modelos de regressão denominada de modelos lineares generalizados (GLM). Nos GLM a ideia básica é a adaptação da regressão linear ordinária a diferentes tipos de dados, isto é, abrir o 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 o valor esperado da variável resposta e o preditor linear (SSI, 2011). Os modelos lineares generalizados são uma abordagem atrativa porque fornecem uma estrutura teórica geral para muitos dos modelos estatísticos comumente encontrados e também simplificam a implementação desses modelos em diferentes *softwares* estatísticos, uma vez que, essencialmente, os mesmos algoritmos podem ser utilizados para a estimação, inferência e avaliação da adequação do modelo para todos os GLM (JACKMAN, 2011).

Neste capítulo serão abordados o conceito de família de distribuição exponencial e suas propriedades, a estrutura dos GLM, com estimação de parâmetros e algoritmo de estimação.

1.1 Família exponencial de distribuições uniparamétricas

Um conceito importante no modelo logístico é o de família exponencial de distribuição.

Segundo definição e notação constante em Cordeiro e Demétrio (2007), a distribuição da variável aleatória *X* pertence à família exponencial uniparamétrica se sua função de probabilidade ou função de densidade de probabilidade pode ser escrita na forma:

$$f(x;\theta) = h(x)\exp\{\eta(\theta)t(x) - b(\theta)\} \quad (1.1)$$

em que as funções $\eta(\theta)$, $b(\theta)$, $t(x)$ e $h(x)$ assumem valores em subconjunto dos reais, com suporte $A = \{x: f(x,\theta) > 0\}$ não dependente de θ.

As funções $\eta(\theta)$, $b(\theta)$ e $t(x)$ não são únicas. Assim, fazendo $\eta(\theta) = \theta$ e $t(x) = x$ em (1.1), obtém-se a forma canônica da família exponencial uniparamétrica:

$$f(x;\theta) = h(x)\exp\{\theta x - b(\theta)\} \quad (1.2)$$

onde θ é chamado de parâmetro canônico.

A expressão (1.2) ainda pode ser reescrita para:

$$f(x;\theta) = \exp\{\theta x - b(\theta) + ln[h(x)]\}. \quad (1.3)$$

Alguns resultados importantes da família exponencial uniparamétrica:

a) logaritmo da função de verossimilhança para uma única observação:

$$l(x;\theta) = \theta x - b(\theta) + ln[h(x)]$$

b) função escore: *U*

$$U(\theta) = \frac{dl(x;\theta)}{d\theta}$$

$$U(\theta) = x - b'(\theta)$$

São propriedades da função escore: E(U) = 0 e Var(U) = –E[$d^2l(\theta)/d\theta^2$].

c) Informação de Fisher: I_F

$$I_F(\theta) = \text{Var(U)}$$

$$I_F(\theta) = -\mathrm{E}\left[\frac{d^2 l(x;\theta)}{d\theta^2}\right] = \mathrm{E}\left[\left(\frac{dl(x;\theta)}{d\theta}\right)^2\right]$$

d) Função geradora de cumulantes: $b(\theta)$

Os momentos consistem em uma série de medidas (média, variância, assimetria etc.) que caracterizam uma distribuição de uma variável aleatória, e são obtidos por meio da função geradora de momentos. Desta forma, a função geradora de momentos representa outra forma de especificar a distribuição de uma variável aleatória.

Os cumulantes também são medidas que caracterizam uma distribuição e são obtidos por meio da função geradora de cumulantes. Representam uma alternativa aos momentos, pois se duas distribuições apresentam momentos idênticos, então terão cumulantes idênticos e vice-versa. O uso da função geradora de cumulantes como alternativa para a geradora de momentos pode simplificar alguns tratamentos teóricos.

Especificamente, no caso da família exponencial uniparamétrica, tem-se que a função geradora de cumulantes é a própria $b(\theta)$ (CORDEIRO; DEMÉTRIO, 2007) e

$$\mathrm{E}(X) = \mu = b'(\theta)$$
$$\mathrm{Var}(X) = b''(\theta)$$

e) Estatística suficiente:

Supondo que X_1, X_2, …, X_n sejam n variáveis aleatórias independentes e identicamente distribuídas, a distribuição conjunta de X_1, …, X_n , usando (1.1), é dada por:

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

$$f(x_1, \dots, x_n; \theta) = \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left\{\eta(\theta) \sum_{i=1}^{n} t(x_i) - nb(\theta)\right\}.$$

Pelo critério da fatoração de Neyman-Fisher, tem-se que $\sum_{i=1}^{n} T(X_i)$ é estatística suficiente para θ.

1.2 Família exponencial de distribuições biparamétricas

À família exponencial uniparamétrica, Nelder e Wedderburn incorporaram distribuições biparamétricas através da introdução de um parâmetro de perturbação $\phi > 0$, redefinindo a família exponencial em (1.3) (CORDEIRO; DEMÉTRIO, 2007; (RODRÍGUEZ, 2007):

$$f(x;\theta,\phi) = \exp\left\{\frac{1}{a(\phi)}[x\theta - b(\theta)] + c(x,\phi)\right\},$$

sendo θ o parâmetro canônico, $\phi > 0$ o parâmetro de escala conhecido e $b(\cdot)$ e $c(\cdot)$ funções conhecidas. Em geral, $a(\phi) = \phi/p$, sendo p o peso *a priori*.

Alguns resultados importantes da família exponencial biparamétrica:

a) logaritmo da função de verossimilhança para uma única observação:

$$l(x;\theta,\phi) = \frac{1}{a(\phi)}[x\theta - b(\theta)] + c(x,\phi)$$

b) função escore: *U*

$$U(\theta) = \frac{dl(x;\theta,\phi)}{d\theta}$$

$$U(\theta) = \frac{1}{a(\phi)}[x - b'(\theta)]$$

c) Informação de Fisher: I_F

$$I_F(\theta) = \mathrm{Var(U)}$$

$$I_F(\theta) = -\mathrm{E}\left[\frac{d^2 l(x;\theta,\phi)}{d\theta^2}\right] = \mathrm{E}\left[\left(\frac{dl(x;\theta)}{d\theta}\right)^2\right]$$

d) Função geradora de cumulantes: $b(\theta)$

$$\mathrm{E}(X) = \mu = \mathrm{b}'(\theta)$$

$$\mathrm{Var}(X) = a(\phi)b''(\theta)$$

e) Estatística suficiente:

Supondo que $X_1, X_2, \dots, X_n$ sejam n variáveis aleatórias independentes e identicamente distribuídas. A distribuição conjunta de $X_1, \dots, X_n$ é dada por:

$$f(x_1, \dots, x_n; \theta, \phi) = \exp\left\{\sum_{i=1}^{n} c(x_i, \phi)\right\} \cdot \exp\left\{\frac{1}{a(\phi)}\left[\theta \sum_{i=1}^{n} x_i - nb(\theta)\right]\right\}.$$

Pelo critério da fatoração de Neyman-Fisher, tem-se que $\sum_{i=1}^{n} X_i$ é estatística suficiente para .

1.3 Estrutura dos modelos lineares generalizados

No contexto dos modelos de efeito fixo, os GLM envolvem uma única variável resposta Y associada a um conjunto de variáveis explicativas $x_1, x_2, \dots, x_p$, e uma amostra aleatória de n observações independentes. Esses modelos apresentam três componentes:

(i) Componente aleatório

O componente aleatório é formado por um conjunto de variáveis aleatórias independentes $Y_1, Y_2, \dots, Y_n$ provenientes de uma mesma distribuição pertencente à família exponencial na forma canônica, com médias $\mu_1, \mu_2, \dots, \mu_n$:

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\} \qquad (1.4)$$

em que $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ e $\mathrm{Var}(Y_i) = a(\phi)b''(\theta_i)$.

***(ii) Componente sistemático*:**

O componente sistemático é formado pelas variáveis explicativas, as quais entram na forma de modelo linear, dando origem ao vetor de preditores lineares:

$$\boldsymbol{\eta} = \boldsymbol{X\beta}, \qquad \text{onde } \eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \qquad (1.5)$$

sendo $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ a matriz do modelo e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ o vetor de parâmetros.

***(iii) Função de ligação*:**

A função de ligação é a função que relaciona os componentes aleatório e sistemático, isto é, relaciona a média ao preditor linear:

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

sendo $g(\cdot)$ uma função monótona e diferenciável.

Quando uma função de ligação resulta no preditor linear da mesma forma que no parâmetro canônico, isto é, $\eta_i = g(\mu_i) = \theta_i$, então a ligação é dita *ligação canônica* (RODRÍGUEZ, 2007). Tais ligações simplificam o algoritmo de estimação dos parâmetros de interesse.

Vale ressaltar que os parâmetros θ_i da família exponencial de distribuição não são de interesse direto na especificação do modelo, mas sim um conjunto de parâmetros $\beta_1, \ldots, \beta_p$ cuja combinação linear é igual a alguma função do valor esperado de Y_i.

Cordeiro e Demétrio (2007), Fox (2008) e Paula (2010) concordam que a escolha do GLM adequado depende da definição destes três componentes. A escolha da distribuição do componente aleatório está vinculada à natureza dos dados e ao seu intervalo de variação; a matriz do modelo, $\boldsymbol{X}$, deve ser de posto completo; e a função de ligação depende do problema em particular.

Cordeiro e Demétrio (2007) observam que, ao contrário do modelo de regressão clássico, nos GLM a variável resposta não é obtida a partir da aditividade entre a média μ e o erro aleatório ε. Portanto, define-se uma distribuição para a variável resposta que representa os dados, e não uma distribuição para o erro aleatório.

1.4 Estimação dos parâmetros dos GLM

Há vários métodos para a estimação dos parâmetros dos GLM. Neste trabalho será apresentado o método de máxima verossimilhança, que possui propriedades ótimas como consistência e eficiência assintótica.

A estimativa de máxima verossimilhança para o vetor $\boldsymbol{\beta}$ envolve a resolução de um sistema de equações normalmente não linear. Assim, métodos numéricos se fazem

necessários. O método aqui desenvolvido será o método numérico de Newton-Raphson, com uma modificação feita por Fisher. Tal modificação provém de um algoritmo proposto por Fisher para o modelo probito, sendo posteriormente incorporada aos modelos logísticos e log-linear. Usada nos GLM, essa modificação acaba por simplificar o método de Newton-Raphson (AGRESTI,2007).

Cordeiro e Demétrio (2007) descrevem o método da estimativa de máxima verossimilhança (EMV), dado por Nelder e Wedderman (1972), de forma muito didática, a qual será aqui reproduzida, com acréscimo de passagens não abordadas.

1.4.1 Método numérico de Newton-Raphson

Originalmente, o método Newton-Raphson (MNR) foi desenvolvido para determinar raízes de funções por processos numéricos. Ao longo dos anos, adaptações foram feitas para atender a outras necessidades, como a maximização da função de verossimilhança, que procura determinar x de tal forma que $f'(x) = 0$, isto é, procura determinar as raízes da primeira derivada de uma função.

O MNR é um processo iterativo baseado numa aproximação da série de Taylor em torno do ponto x_0:

$$f(x) = \sum_{n=0}^{\infty} \frac{(x - x_0)^n f^{(n)}(x_0)}{n!}, \text{ com } f^{(0)}(x_0) = f(x_0).$$

Assim, para uma aproximação até a primeira derivada (n = 0, 1) a aproximação para $f(x)$ é dada por:

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0).$$

Desejando-se determinar x tal que $f(x) = 0$, basta fazer:

$$f(x_0) + (x - x_0)f'(x_0) \approx 0$$

$$x \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

No processo iterativo, o valor de x encontrado num passo torna-se o novo ponto (um novo x_0) em torno do qual se desenvolverá uma nova aproximação para x. E assim sucessivamente. De uma forma mais geral,

$$x^{(m+1)} = x^{(m)} - \frac{f\left(x^{(m)}\right)}{f'^{(x^{(m)})}}$$

$$x^{(m+1)} = x^{(m)} - \left[f'\left(x^{(m)}\right)\right]^{-1} \cdot f\left(x^{(m)}\right), \qquad (1.6)$$

sendo $x^{(m+1)}$ o valor de x no passo $(m + 1)$, $x^{(m)}$ o valor de x no passo (m) e $f\left(x^{(m)}\right)$ e $f'\left(x^{(m)}\right)$ a função $f(x)$ e sua derivada avaliadas em $x^{(m)}$. Esse processo é repetido até que se obtenha a convergência.

A demonstração da interpretação geométrica do MNR, encontrada em Santos (1982), está baseada na Figura 1.

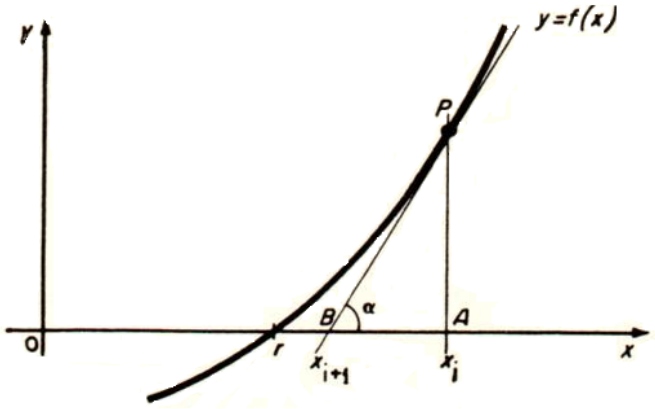


Figura 1 – Interpretação gráfica do método de Newton-Raphson
Fonte: Santos, 1982, p.63.

No triângulo retângulo ABP, A é a localização de x_i no eixo $\overrightarrow{Ox}$, $P = (x_i, f(x_i))$ e B é o ponto x_{i+1} obtido pela interseção da reta tangente à curva no ponto P com $\overrightarrow{Ox}$. Portanto,

$$\text{tg}\,\alpha = \frac{f(x_i)}{x_i - x_{i+1}}.$$

Por definição, $\text{tg}\,\alpha = f'(x_i)$. Portanto,

$$f'(x_i) = \frac{f(x_i)}{x_i - x_{i+1}}$$

$$x_i - x_{i+1} = \frac{f(x_i)}{f'(x_i)}$$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)},$$

que é a fórmula de recorrência do método, onde x_i corresponde a $x^{(m)}$ e x_{i+1} a $x^{(m+1)}$.

A Figura 2 mostra geometricamente o MNR para várias iterações.

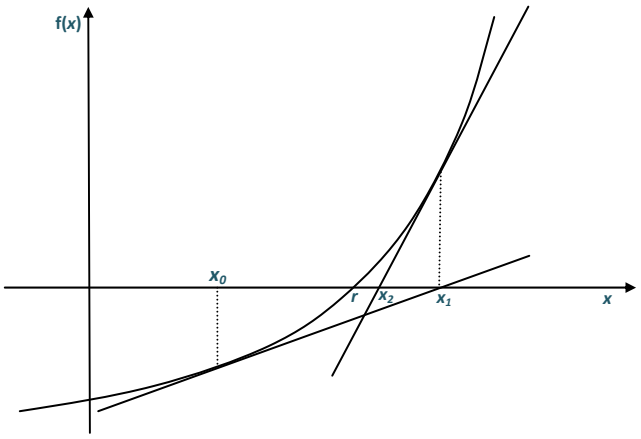


Figura 2 – Interpretação gráfica do método de Newton-Raphson para várias iterações
Fonte: Ruggiero e Lopes, 2006, p. 68.

Um dos critérios de convergência é dado pelo erro relativo entre duas iterações consecutivas:

$$\left|\frac{x^{(m+1)} - x^{(m)}}{x^{(m)}}\right| < \xi, \tag{1.7}$$

para ξ suficientemente pequeno.

A convergência ocorre rapidamente, pois é quadrática. Porém algumas condições devem ser satisfeitas para que possa ser garantida:

(i) é preciso que $f(x)$, $f'(x)$ e $f''(x)$ sejam contínuas num intervalo aberto que contém a raiz;

(ii) é preciso que, em cada passo, $f'\left(x^{(m)}\right)$ seja diferente de zero.

De maneira geral, afirmam Ruggiero e Lopes (1996), o MNR converge desde que x_0 seja tomado "suficientemente próximo" da raiz.

1.4.2 Estimador de máxima verossimilhança

Seja *f* uma função de distribuição de probabilidade da família exponencial, com parâmetro de dispersão ϕ conhecido e $a(\phi) = \phi$:

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{1}{\phi}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\},$$

em que

$$\mathrm{E}(Y_i) = \mu_i = b'(\theta_i) \tag{1.8}$$

$$\frac{d\mu_i}{d\theta_i} = b''(\theta_i) = V_i = V(\mu_i) \quad \text{(função de variância)}$$

$$\text{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i$$

$$\frac{d\theta_i}{d\mu_i} = V_i^{-1} \Longrightarrow \theta_i = \int V_i^{-1} d\mu_i = q(\mu_i). \tag{1.9}$$

Da relação entre o componente sistemático e a função de ligação dos GLM, deduz-se que:

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \tag{1.10}$$

$$\eta_i = g(\mu_i) \Longrightarrow g^{-1}(\eta_i) = \mu_i. \tag{1.11}$$

Pode-se, então, estabelecer a seguinte sequência:

$$\theta_i = q(\mu_i),\ \ \mu_i = g^{-1}(\eta_i),\ \ \eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}. \tag{1.12}$$

Desta forma, dado o vetor $\boldsymbol{y}$, é possível definir o logaritmo neperiano da função de verossimilhança, isto é, a log-verossimilhança, $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ como $\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \ell(\boldsymbol{\beta})$, a partir da composição de funções:

$$\ell(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi). \tag{1.13}$$

Por consequência, o vetor escore fica definido em função das derivadas parciais de primeira ordem da função $\ell(\boldsymbol{\beta})$, com dimensão *p*:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Usando a regra da cadeia para derivadas e observando a sequência de funções dada em (1.12), o vetor escore possui elemento típico

$$U_r(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_r} = \frac{1}{\phi}\sum_{i=1}^{n} \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r}.$$

A partir de (1.13) e (1.8), de (1.9) e de (1.10), reescreve-se $U_r(\boldsymbol{\beta})$:

$$U_r(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V_i}\frac{d\mu_i}{d\eta_i}x_{ir}\,, \qquad r = 1, \ldots, p \tag{1.14}$$

O sistema de equações $U_r(\boldsymbol{\beta}) = 0$, $r = 1, \ldots, p$ leva a um estimador de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\beta}$. Como normalmente suas equações são não lineares, a estimação é feita pelo método numérico de Newton-Raphson, usando a modificação de Fisher, chamada de método escore de Fisher.

Para usar o MNR será preciso estabelecer algumas analogias com a expressão (1.6). Primeiro é preciso lembrar que o que se deseja determinar é o vetor $\widehat{\boldsymbol{\beta}}$ tal que a

log-verossimilhança seja máxima, isto é, um vetor $\widehat{\boldsymbol{\beta}}$ de tal forma que as derivadas parciais de primeira ordem da função log-verossimilhança sejam identicamente nulas, $\boldsymbol{U}(\boldsymbol{\beta}) \equiv \mathbf{0}$. Assim, à função f far-se-á corresponder a matriz de derivadas parciais de primeira ordem da função log-verossimilhança, $\boldsymbol{U}$; à primeira derivada de f corresponderá à matriz de derivadas parciais de segunda ordem da função log-verossimilhança; à variável x, o vetor $\boldsymbol{\beta}$. Assim procedendo, tem-se que

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left[-\left(\mathbf{J}^{(m)}\right)^{-1}\right]\mathbf{U}^{(m)}$$

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left(\mathbf{J}^{(m)}\right)^{-1}\mathbf{U}^{(m)}, \qquad (1.15)$$

sendo $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\beta}^{(m+1)}$ os vetores de parâmetros estimados nos passos (m) e (m +1), $\mathbf{U}^{(m)}$ o vetor escore avaliado no passo (m) e $\left(\mathbf{J}^{(m)}\right)^{-1}$ a inversa da negativa da matriz de derivadas parciais de segunda ordem de $\ell(\boldsymbol{\beta})$ (todas as derivadas de segunda ordem obtidas a partir das derivadas de primeira ordem que compõem o vetor escore $\boldsymbol{U}$), avaliada no passo (m). A matriz **J** é a matriz de informação observada, com elemento típico

$$\mathrm{J}_{rs} = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s}.$$

Nem sempre estas derivadas são facilmente avaliadas. No caso dos GLM usa-se o método escore de Fisher, que em geral é mais simples, substituindo-se a matriz de informação observada, **J**, pela matriz de informação esperada de Fisher, **K**, com elemento típico

$$\mathrm{K}_{rs} = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s}\right] = \mathrm{E}\left[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_r}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_s}\right]$$

$$\mathrm{K}_{rs} = \frac{1}{\phi^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\mathrm{E}\left[(y_i - \mu_i)(y_j - \mu_j)\right]\frac{1}{V_i}\frac{1}{V_j}\frac{d\mu_i}{d\eta_i}\frac{d\mu_j}{d\eta_j}x_{ir}x_{js}.$$

Se $i \neq j$, tem-se que $\mathrm{E}\left[(y_i - \mu_i)(y_j - \mu_j)\right] = \mathrm{Cov}(Y_i, Y_j) = 0$, pois Y_i e Y_j são independentes. Se $i = j$, $\mathrm{E}[(y_i - \mu_i)^2] = \mathrm{Var}(Y_i) = \phi V_i$. Assim, K_{rs} fica reduzido a:

$$\mathrm{K}_{rs} = \frac{1}{\phi^2}\sum_{i=1}^{n}\phi V_i\left(\frac{1}{V_i}\right)^2\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ir}x_{is}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\frac{1}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ir}x_{is}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n} w_i x_{ir}x_{is},$$

sendo $w_i = \frac{1}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2$ denominado peso (1.16)

Portanto, a matriz **K** pode ser escrita como

$$\mathbf{K} = \phi^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}, \tag{1.17}$$

sendo $\mathbf{W} = \mathrm{diag}\{w_1, w_2, \ldots, w_n\}$ uma matriz diagonal de pesos que traz a informação sobre a distribuição e a função de ligação usadas e poderá incluir também um termo para peso a priori. No caso de ligações canônicas tem-se que $w_i = V_i$, pois $V_i = V(\mu_i) = d\mu_i/d\eta_i$.

Voltando à expressão (1.15), com a substituição da matriz **J** pela matriz **K**, obtém-se:

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} + \left(\mathbf{K}^{(m)}\right)^{-1}\mathbf{U}^{(m)} \\ \mathbf{K}^{(m)}\boldsymbol{\beta}^{(m+1)} &= \mathbf{K}^{(m)}\boldsymbol{\beta}^{(m)} + \mathbf{U}^{(m)}. \end{aligned} \tag{1.18}$$

Multiplicando e dividindo por $d\mu_i/d\eta_i$ a expressão de $U_r(\boldsymbol{\beta})$ em (1.14) obtém-se:

$$U_r(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2\frac{d\eta_i}{d\mu_i}x_{ir}.$$

Logo,

$$\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{U} = \frac{1}{\phi}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{G}(\boldsymbol{y} - \boldsymbol{\mu}),$$

onde $\mathbf{G} = \mathrm{diag}\{d\eta_1/d\mu_1, \ldots, d\eta_n/d\mu_n\} = \mathrm{diag}\{g'(\mu_1), \ldots, g'(\mu_n)\}$, sendo g a função de ligação.

Substituindo as formas matriciais de **K** e **U** em (1.18), obtém-se:

$$\begin{aligned} \mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m+1)} &= \mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m)} + \mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{G}^{(m)}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= \mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\left[\mathbf{X}\boldsymbol{\beta}^{(m)} + \mathbf{G}^{(m)}(\boldsymbol{y} - \boldsymbol{\mu})\right] \\ &= \mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\left[\boldsymbol{\eta}^{(m)} + \mathbf{G}^{(m)}(\boldsymbol{y} - \boldsymbol{\mu})\right]. \end{aligned}$$

A partir daí, define-se a variável dependente ajustada $\mathbf{z} = \boldsymbol{\eta} + \mathbf{G}(\boldsymbol{y} - \boldsymbol{\mu})$. Logo,

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)}. \tag{1.19}$$

Cordeiro e Demétrio (2007) resumem o resultado acima, lembrando que:

(i) a equação matricial (1.19) é válida para qualquer GLM;

(ii) não depende do parâmetro de dispersão ϕ;

(iii) essa equação equivale a calcular repetidamente uma regressão linear ponderada (IWLS) de uma variável dependente ajustada **z** sobre a matriz **X**, usando uma função de peso **W** que se modifica durante o processo iterativo;

(iv) as funções de variância e de ligação entram no processo iterativo através de **W** e de **z**;

(v) os z_i ajustados são não correlacionados, pois Cov(**z**) = **G** Cov(**Y**) **G** = $\phi \boldsymbol{W}^{-1}$, que é uma matriz diagonal.

1.5 Algoritmo de estimação

(i) Estimativa inicial para z

Cada observação y_i pode ser considerada uma estimativa para seu valor médio μ_i. A partir de (1.10) e (1.11) e de (1.16) pode-se escrever:

$$\eta_i{}^{(1)} = g\left(\mu_i{}^{(1)}\right) = g(y_i) \quad \text{e}$$

$$w_i{}^{(1)} = \frac{1}{V(y_i)[g'(y_i)]^2}.$$

Para $m = 1$, faz-se $\boldsymbol{z}^{(1)} = \boldsymbol{\eta}^{(1)}$, obtendo-se $\boldsymbol{\beta}^{(2)}$,

$$\boldsymbol{\beta}^{(2)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(1)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(1)}\boldsymbol{\eta}^{(1)}.$$

Se a função g não for definida para alguns valores de y_i, o processo não poderá ser iniciado sem fazer os ajustes necessários, substituindo $\boldsymbol{y}$ por $(\boldsymbol{y} + \boldsymbol{c})$ tal que $\mathrm{E}[g(\boldsymbol{y} + c)]$ seja o mais próximo possível de $g(\boldsymbol{\mu})$. Detalhes a respeito podem ser vistos em Cordeiro e Demétrio (2007).

(ii) Cálculo de $\boldsymbol{\beta}^{(\boldsymbol{m+1})}$

Para $m \geq 2$, utiliza-se o cálculo iterativo apresentado em (1.19), a partir dos passos:

(1º) Obtenção das estimativas de $\boldsymbol{\eta}$ e $\boldsymbol{\mu}$

$$\eta_i{}^{(m)} = \sum_{r=1}^{p} x_{ir}\beta_r{}^{(m)},$$

$$\mu_i{}^{(m)} = g^{-1}\left(\eta_i{}^{(m)}\right).$$

(2º) Obtenção de **z** e **W**

$$z_i^{(m)} = \eta_i^{(m)} + g'\left(\mu_i^{(m)}\right)\left(y - \mu_i^{(m)}\right),$$

$$w_i{}^{(m)} = \frac{1}{V\left(\mu_i^{(m)}\right)\left[g'\left(\mu_i^{(m)}\right)\right]^2}.$$

(3º) Obtenção da estimativa de $\boldsymbol{\beta}$

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)}.$$

(iii) Verificação de convergência

Fazendo analogia com o erro relativo apresentado em (1.7), um dos critérios de convergência para multivariada poderia ser:

$$\sum_{r=1}^{p}\left(\frac{\beta_r^{(m+1)} - \beta_r^{(m)}}{\beta_r^{(m)}}\right)^2 < \xi,$$

tomando-se para ξ um valor suficientemente pequeno.

1.6 Interpretações subjacentes ao método de Newton-Raphson

Para melhor compreender as consequências do resultado (1.19) é preciso entender o que a própria fórmula traduz e o significado do vetor escore, da matriz de pesos e da matriz de informação de Fisher.

Agresti (2007) trata a EMV como uma estimativa de quadrados mínimos (EMQ): "cada ciclo no MNR representa um tipo de ajuste de quadrados mínimos ponderados. Isto é, uma generalização dos quadrados mínimos ordinários que leva em consideração a não constância da variância". Esta afirmação é justificada pela sequência de raciocínios dada a seguir.

Em modelos lineares ordinários de posto completo (REGAZZI, 2010) a estimativa de quadrados mínimos (EMQ) para o vetor $\boldsymbol{\beta}$ é dada por:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \qquad (1.20)$$

Comparando a EMV dada em (1.19) com a expressão acima, verifica-se que: $\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}$ corresponde a $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$, porém apresentando uma ponderação pelo

peso; $\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)}$ corresponde a $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ ponderado, sendo a variável dependente $\boldsymbol{Y}$ substituída por **z**, que é a variável dependente ajustada. Logo, a estimativa de máxima verossimilhança para os GLM é uma estimativa de quadrados mínimos.

O vetor escore **U**, por ser constituído de todas as derivadas parciais de primeira ordem, ao ser igualado ao vetor nulo leva à pesquisa dos pontos críticos da função de verossimilhança $\ell(\boldsymbol{\beta})$. A existência de ponto crítico indica não somente um máximo local, mas um máximo absoluto. Mas é preciso que exista esse ponto crítico. A Figura 3 mostra duas curvas com concavidade para baixo, sendo que a primeira tem ponto de máximo e a segunda não.

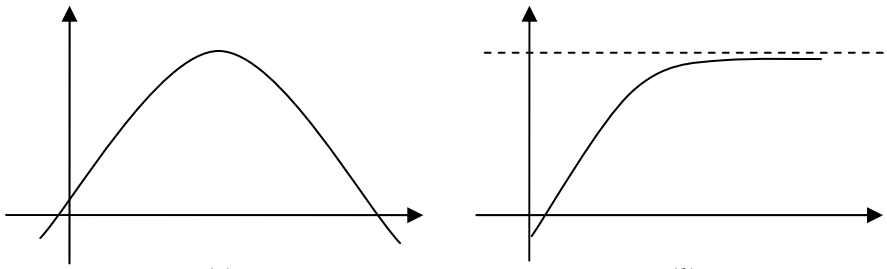


Figura 3 – Funções com concavidade para baixo: (*a*) com ponto de máximo e (*b*) sem ponto de máximo

Agresti (2007) mostra que a matriz de pesos **W**, por sua vez, tem na composição de seu elemento característico o inverso da variância. Isto quer dizer que quanto menor a variância de uma observação, maior será seu peso na determinação da estimativa do vetor $\boldsymbol{\beta}$. Outra observação importante é que os pesos são recalculados em cada iteração, levando em consideração a não constância de variâncias (ausência de homocedasticidade) da variável resposta.

Quanto à matriz de informação de Fisher, **K**, esta está associada à curvatura da função log-verossimilhança. Quanto maior a curvatura, isto é, quanto mais fechada a curva, maior a informação sobre o valor do parâmetro. A matriz de informação de Fisher fornece o desvio padrão de $\widehat{\boldsymbol{\beta}}$, através da raiz quadrada dos elementos da diagonal de sua inversa. Logo, quanto maior a curvatura, menor o desvio padrão. Uma grande curvatura implica que o logaritmo da função de verossimilhança cai rapidamente à medida que $\boldsymbol{\beta}$ se afasta de $\widehat{\boldsymbol{\beta}}$ (AGRESTI, 2007). Vale observar que um dos fatores componentes da matriz **K** é a matriz de pesos **W**. Esta matriz diagonal é não negativa. Quando positiva-definida, ela garante a existência de mínimos, mas se positiva-

semidefinida, a matriz **K** apresentará em sua diagonal pelo menos um elemento igual a zero, levando a uma variância infinita para algum componente do vetor estimado $\widehat{\boldsymbol{\beta}}$.

1.7 Covariância e intervalo de confiança

Como dito no item anterior, a informação de Fisher está associada ao desvio padrão de $\widehat{\boldsymbol{\beta}}$. Assim,

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}) = \phi\left(\boldsymbol{X}^{\mathrm{T}}\widehat{\mathbf{W}}\boldsymbol{X}\right)^{-1},$$

onde $\widehat{\mathbf{W}}$ é a matriz de pesos avaliada em $\widehat{\boldsymbol{\beta}}$, isto é, a matriz de covariância de $\widehat{\boldsymbol{\beta}}$ é o inverso da matriz de informação de Fisher avaliada em $\widehat{\boldsymbol{\beta}}$.

A matriz $\phi\left(\boldsymbol{X}^{\mathrm{T}}\widehat{\mathbf{W}}\boldsymbol{X}\right)^{-1}$ é também a matriz de covariância assintótica de $\widehat{\boldsymbol{\beta}}$, então a Var($\widehat{\boldsymbol{\beta}}_r$) é o elemento ($r$, r) dessa matriz. O intervalo de confiança a 100(1 – α)% para β_r pode ser obtido por:

$$\mathrm{IC}(\beta_r)_{1-\alpha} : \hat{\beta}_r \pm Z_{\frac{\alpha}{2}}\left[\mathrm{Var}\left(\hat{\beta}_r\right)\right]^{1/2}.$$

CAPÍTULO 2 – MODELO LOGÍSTICO

Conhecido desde os anos 50 e popularizado por Cox em 1970, o modelo logístico ou modelo de regressão logística constitui um tipo especial de modelo linear generalizado – cujos aspectos teóricos básicos foram vistos no capítulo 1 – desenvolvido para o caso em que a variável resposta é categórica binária. Através dele pode-se avaliar o impacto que as variáveis explicativas exercem sobre a variável resposta (GARSON, 2010).

A facilidade de interpretação dos parâmetros fez com que o modelo logístico se tornasse um dos principais métodos de modelagem estatística. Isto é notado no fato de que até mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns pesquisadores têm dicotomizado a resposta de modo que a probabilidade de sucesso possa ser modelada através de regressão logística, como acontece em análise de sobrevivência discreta (PAULA, 2004).

Neste capítulo serão introduzidos os conceitos necessários à compreensão de modelos binários e, em especial, o modelo de regressão logística. Para este será feita a adaptação do algoritmo de estimação bem como dos resultados obtidos para os GLM.

2.1 Dados binários

Em várias situações experimentais aparece a necessidade de trabalhar com variáveis discretas, cujos valores são categorizados, fugindo ao padrão de continuidade. Como exemplo, pode-se citar: a resposta a um tratamento de cisto no ovário foi “excelente”, “boa”, “razoável” ou “ruim”; um experimento envolvendo vários tipos de solo pode ter como resposta “houve germinação” ou “não houve germinação” da semente; uma pessoa pode pertencer a um grupo religioso “católico”, “protestante”, “evangélico”, “espírita”, “ateu”, “outros”; o tempo de falha de um equipamento pode ser “menor que 3 anos”, “entre 3 e 5 anos”, “maior que 5 anos”, o número de acidentes ocorridos num determinado período de tempo pode ser “0”, “1”, “2”, “3”,… Esse tipo de variável é chamada de categórica.

Segundo Agresti (2007), variável categórica é uma variável qualitativa que tem uma escala de medidas definida em um conjunto de categorias. Escalas categóricas são

muito difundidas em ciências sociais e ciências médicas, mas também aparecem em psicologia, epidemiologia e saúde pública, educação, zoologia, agronomia, controle de qualidade industrial, marketing e muitas outras áreas. Essas variáveis podem aparecer num experimento tanto como variáveis explicativas quanto como variáveis respostas. Entende-se por variáveis explicativas, explanatórias, preditoras ou covariáveis as variáveis independentes, normalmente chamadas de variável X; e por variáveis respostas, as variáveis dependentes, normalmente chamadas de variável Y.

Agresti (2002) fala em dois tipos primários de variáveis categóricas: a ordinal e a nominal. Porém, apresenta também um terceiro tipo: as variáveis intervalares.

No presente trabalho, as variáveis respostas serão do tipo categórico binário (ou dicotômico) – caso especial das variáveis categóricas ordinais – as quais podem apresentar apenas dois valores, representados genericamente por "fracasso" ou "sucesso" e sendo expressos na forma ordinal por 0 ou 1. Alguns exemplos:

a) estudo da sobrevivência de enxertos de ameixeiras sob determinadas condições técnicas ("sobrevive" ou "não sobrevive");

b) estudo da "presença" ou "ausência" de um departamento de relação industrial numa firma, de acordo com o seu tamanho;

c) estudo sobre a reincidência de determinado tipo de câncer após tratamento com quimioterapia, num espaço de 5 anos ("houve reincidência", "não houve reincidência");

d) estudo sobre a necessidade de assistência técnica a um aparelho antes do término da garantia ("necessária", "não necessária").

2.2 Modelos para dados binários

Agresti (2002) faz a ponderação a seguir sobre os modelos para dados binários.

Considere-se um experimento em que a variável resposta Y é binária, isto é, $Y = 1$ para "sucesso" ou $Y = 0$ para "fracasso". Seja $X = x$ a variável preditora à qual se associa um dos valores que Y pode assumir com as seguintes probabilidades:

$$\mathrm{P}(Y = 1 | X = x) = \pi(x) \text{ e } \mathrm{P}(Y = 0 | X = x) = 1 - \pi(x).$$

Isto é, a probabilidade de Y assumir qualquer um dos valores depende do valor que X assumir. Tem-se também que $\mathrm{E}(Y) = \pi(x)$ e $\mathrm{Var}(Y) = \pi(x)\big(1 - \pi(x)\big)$. Um fato importante a observar é a ausência de homocedasticidade, visto que a variância de Y não é constante em todo o experimento, pois depende de cada valor que X pode apresentar.

Na regressão linear ordinária, tem-se que $\mathrm{E}(Y) = \alpha + \beta x$. Analogamente, pode-se fazer:

$$\pi(x) = \alpha + \beta x \tag{2.1}$$

Esta expressão, chamada de modelo linear de probabilidade, apresenta uma relação linear entre a probabilidade de *x* levar ao sucesso e o próprio *x*. Ou ainda, mostra que a variação de uma unidade em *x* provoca uma variação de β unidades na probabilidade de sucesso.

Sabe-se que $0 \leq \pi(x) \leq 1$, mas a função linear dada em (2.1) pode assumir qualquer valor na reta real, predizendo probabilidades menores que zero e maiores que um para algum x suficientemente pequeno ou grande. O modelo pode fazer um ajuste adequado apenas para um intervalo restrito de valores para x, como mostrado na Figura 1.

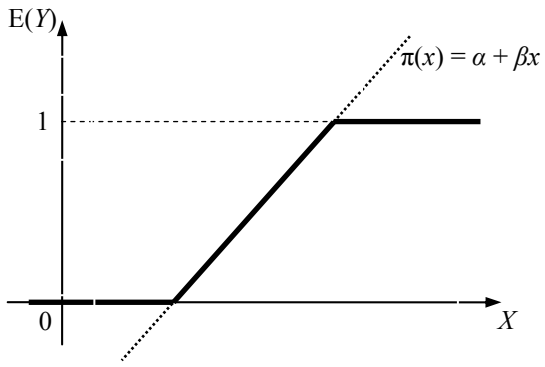


Figura 1 – Gráfico esperado para a probabilidade de sucesso (linha grossa) e gráfico encontrado para $\pi(x) = \alpha + \beta x$
Fonte: Ogliari, [200-], p. 5.

Além disso, a relação entre $\pi(x)$ e x usualmente é não linear. Uma alteração fixa em x pode produzir menos impacto quando a probabilidade está mais próxima de 0 ou de 1, isto é, próximos aos pontos de estabilização, do que quando se encontra no meio desse intervalo (AGRESTI, 2002). Na prática, $\pi(x)$ cresce continuamente ou decresce continuamente conforme x cresce. Assim, a forma sigmoide mostrada na Figura 2 é normalmente mais realista para representar a relação entre x e $\pi(x)$. A questão toda é determinar qual função com forma sigmoide é a mais adequada.

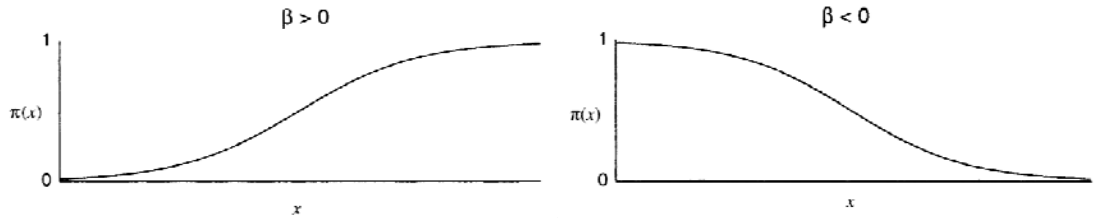


Figura 2 – Sigmoides da função logística
Fonte: Agresti, 2007, p.71.

Cordeiro e Demétrio (2007) apresentam outra visão do problema, a qual permite inferir sobre a função sigmoidal mais adequada. A partir de um ensaio do tipo dose-resposta, definem uma nova variável, conhecida como variável de tolerância ou latente (*threshold*). A dose é a intensidade de um estímulo (inseticida, medicamento, voltagem, força) aplicada a um indivíduo (planta, animal, pessoa, resistor, bloco de concreto); a resposta é binária (morreu/não morreu, melhorou/não melhorou, queimou/não queimou, quebrou/não quebrou). Uma variável de tolerância define um limite a partir do qual a dose aplicada a um indivíduo começa a produzir o efeito desejado, isto é, a probabilidade de sucesso passa a ser 1. Cada indivíduo tem o seu nível de tolerância ou de latência. Seja T a variável aleatória que representa o nível de tolerância. Sua f.d.p. pode ser representada tanto por uma curva simétrica como assimétrica (Figura 3).

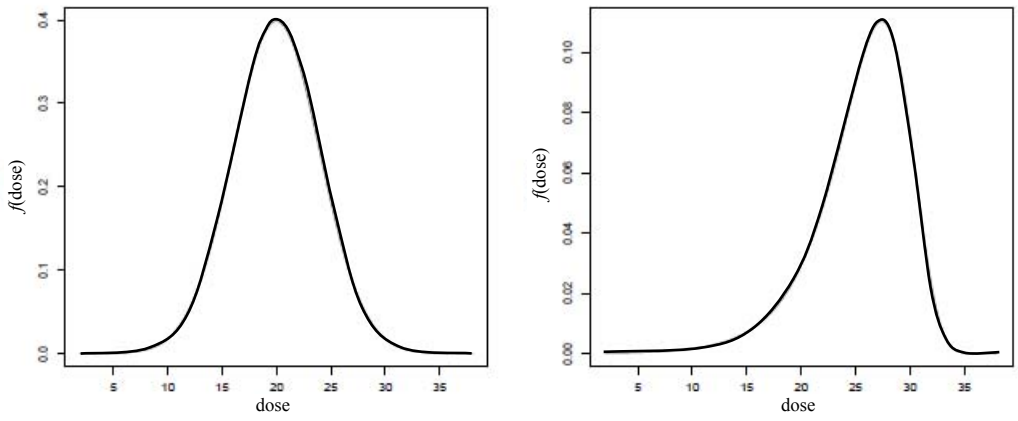


Figura 3 – Dois tipos de distribuições de tolerância: (*a*) simétrica e (*b*) assimétrica
Fonte: Cordeiro e Demétrio, 2007, p. 19.

Num ensaio do tipo dose-resposta tudo que se sabe é que a tolerância de cada indivíduo está acima ou abaixo da dose a ele aplicada. Nestas condições, tem-se que :

$$\begin{cases} Y_i = 1, & \text{sempre que } T_i \leq X_i \\ Y_i = 0, & \text{sempre que } T_i > X_i \end{cases}.$$

Então, para uma dada dose X_i aplicada a um indivíduo ao acaso,

$$\pi_i = \mathrm{P}(Y_i = 1|X_i) = \mathrm{P}(T_i \leq X_i).$$

A $\mathrm{P}(T \leq X)$ é a distribuição de probabilidade acumulada da variável de tolerância de todos os indivíduos da população. A curva da função acumulada de probabilidade tem a forma de uma sigmoide, como mostrada na Figura 4.

O problema passa a ser encontrar uma curva sigmoide que se ajuste bem aos dados. Como se trata de modelos não lineares, mas lineares nos parâmetros, a ideia é que essa curva se transforme numa reta para que procedimentos usuais de regressão possam ser usados para a estimação. As curvas de distribuição de tolerância mais comuns estão apresentadas na Figura 5, com suas respectivas sigmoides. Para cada uma dessas sigmoides há uma função de linearização.

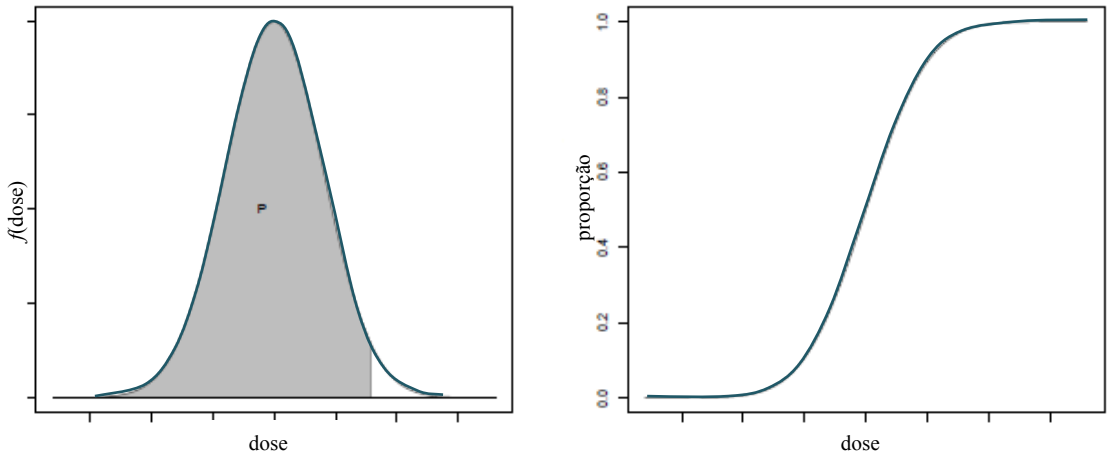


Figura 4 – Área sob a curva de tolerância e correspondente distribuição acumulada
Fonte: Cordeiro e Demétrio, 2007, p.20.

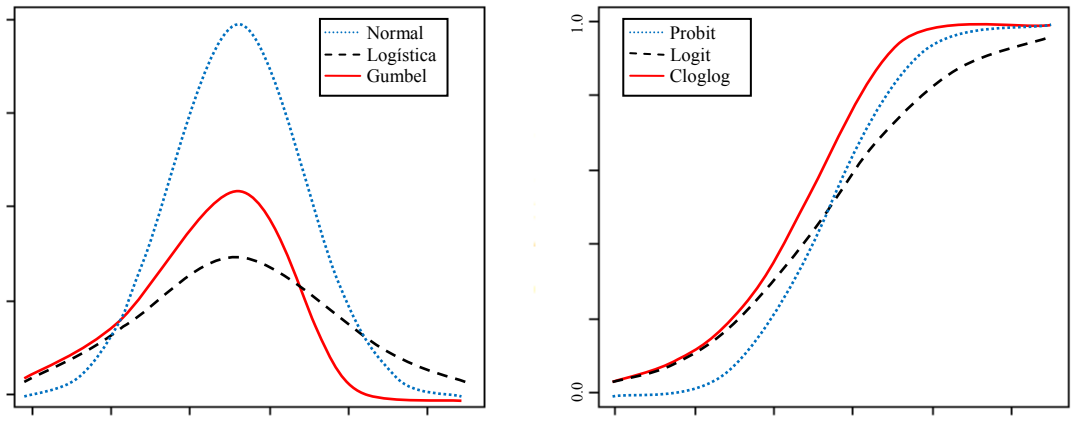


Figura 5 – Distribuições de tolerância e respectivas sigmoides
Fonte: Cordeiro e Demétrio, 2007, p.20.

***(i) Modelo probito** ("**Prob**ability un**it**")*

No modelo probito assume-se que $T \sim N(\mu, \sigma^2)$. Sua f.d.p. é dada por:

$$f_t(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right].$$

Desta forma

$$Z = \frac{T-\mu}{\sigma} \sim N(0,1) \text{ e}$$

$$\pi_i = \mathrm{P}(T \leq x_i) = \mathrm{P}\left(Z \leq -\frac{\mu}{\sigma} + \frac{1}{\sigma} x_i\right) = \mathrm{P}(Z \leq \alpha + \beta x_i),$$

para $\alpha = -\mu/\sigma$ e $\beta = 1/\sigma$. Logo,

$$\pi_i = \Phi(\alpha + \beta x_i),$$

que é uma função não linear em um conjunto linear de parâmetros, onde $\Phi(\cdot)$ representa a função de distribuição normal padrão. A linearização é dada por:

$$\text{probit}\,(\pi_i) = \Phi^{-1}(\pi_i) = \alpha + \beta x_i.$$

***(ii) Modelo logito** ("**Log**istic un**it**")*

A distribuição logística é similar à distribuição normal, com caudas mais longas. Se for assumido que T tem esta distribuição logística com parâmetros $\mu \in \mathbb{R}$ e $\tau > 0$, sua f.d.p. será dada por:

$$f_t(t; \mu, \tau) = \frac{1}{\tau} \cdot \frac{\exp\left(\frac{t-\mu}{\tau}\right)}{\left[1 + \exp\left(\frac{t-\mu}{\tau}\right)\right]^2} = \frac{1}{\tau} \cdot \frac{\exp\left(-\frac{\mu}{\tau} + \frac{1}{\tau} t\right)}{\left[1 + \exp\left(-\frac{\mu}{\tau} + \frac{1}{\tau} t\right)\right]^2},$$

com $\mathrm{E}(T) = \mu$ e $\mathrm{Var}(T) = \pi^2\tau^2/3$. Na expressão acima, fazendo $\alpha = -\mu/\tau$ e $\beta = 1/\tau$, a f.d.p. fica dada por:

$$f_t(t; \alpha, \beta) = \frac{\beta e^{\alpha+\beta t}}{[1 + e^{\alpha+\beta t}]^2}.$$

Sendo $\pi_i = \mathrm{P}(T \leq x_i) = F(x_i)$, tem-se que

$$\pi_i = F(x_i) = \int_{-\infty}^{x_i} \frac{\beta e^{\alpha+\beta t}}{[1 + e^{\alpha+\beta t}]^2}\, dt$$

$$= -\lim_{a \to -\infty} \left(\left[\frac{1}{1 + e^{\alpha+\beta t}} \right]_a^{x_i} \right)$$

$$= -\frac{1}{1 + e^{\alpha+\beta x_i}} + 1$$

Logo,

$$\pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}},$$

que é uma função não linear num conjunto linear de parâmetros, podendo ser linearizada por

$$\text{logit}(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = ln\left(\frac{\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}}{1-\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}}\right)$$

$$\text{logit}(\pi_i) = \alpha + \beta x_i.$$

Apesar dos modelos probito e logito serem similares, o último é mais simples de ser interpretado. Daí a sua popularização.

(iii) Modelo complemento log-log

Se for assumido que a variável T tem distribuição de Gumbel, que é assimétrica, sua f.d.p. será dada por

$$f_t(t;\alpha,\tau) = \frac{1}{\tau}\exp\left(\frac{t-\alpha}{\tau}\right)\exp\left[-\exp\left(\frac{t-\alpha}{\tau}\right)\right], \qquad \alpha \in \mathbb{R}, \tau > 0,$$

com $\text{E}(T) = \alpha + \gamma\tau$ e $\text{Var}(T) = \pi^2\tau^2/6$, sendo $\gamma \approx 0{,}577216$ o número de Euler definido a partir da função digama. Fazendo $\beta_1 = -\alpha/\tau$ e $\beta_2 = 1/\tau$ na expressão acima, resulta que

$$\begin{aligned} f_t(t;\beta_1,\beta_2) &= \beta_2 \exp[\beta_1 + \beta_2 t] \cdot \exp[-\exp(\beta_1 + \beta_2 t)] \\ &= \beta_2 \exp\left[\beta_1 + \beta_2 t - e^{\beta_1+\beta_2 t}\right]. \end{aligned}$$

Sendo $\pi_i = \text{P}(T \le x_i) = F(x_i)$, obtém-se

$$\begin{aligned} \pi_i = F(x_i) &= \int_{-\infty}^{x_i} \beta_2\, e^{\beta_1+\beta_2 t - e^{\beta_1+\beta_2 t}}\, dt \\ &= \lim_{a\to-\infty} \int_a^{x_i} \frac{\beta_2\, e^{\beta_1+\beta_2 t}}{e^{e^{\beta_1+\beta_2 t}}}\, dt \\ &= -\lim_{a\to-\infty} \left(\left[e^{-e^{\beta_1+\beta_2 t}}\right]_a^{x_i}\right) \\ \pi_i &= 1 - e^{-e^{\beta_1+\beta_2 x_i}}, \end{aligned}$$

que é uma função não linear num conjunto linear de parâmetros, podendo ser assim linearizada:

$$\begin{aligned} 1-\pi_i &= e^{-e^{\beta_1+\beta_2 x_i}} \\ -ln(1-\pi_i) &= e^{\beta_1+\beta_2 x_i} \\ ln[-ln(1-\pi_i)] &= \beta_1 + \beta_2 x_i. \end{aligned}$$

2.3 O modelo logístico como GLM

Os modelos logito, probito e complemento log-log são modelos lineares generalizados em que o componente aleatório segue distribuição binomial, as variáveis explicativas entram na forma de uma soma linear de seus efeitos sistemáticos e cada um desses modelos possui uma função de ligação monótona e diferenciável. A demonstração desses fatos será feita apenas para o modelo logístico (modelo logito) ou modelo de regressão logística.

(i) ***Componente aleatório***

Considere-se um experimento cuja realização da variável resposta X seja dicotômica, isto é, 1 para sucesso ou 0 para fracasso. Seja π a probabilidade de sucesso numa única tentativa, Neste caso, $X \sim$ Bernoulli(π), cuja função de probabilidade é dada por:

$$\mathrm{P}(X = x) = \pi^{x}(1-\pi)^{1-x}$$

Se o experimento for executado m vezes, pode-se ter 0, 1, 2, …, m sucessos. O número de sucessos será dado por $y = \sum_{i=1}^{m} x_i$. Desta forma, $Y \sim$ Binomial(m, π), cuja função de probabilidade é dada por:

$$f(y; m, \pi) = \binom{m}{y}\pi^{y}(1-\pi)^{m-y}, \text{com } \pi \in [0,1] \text{ e } A = \{0, 1, 2, 3, \dots, m\}.$$

Demétrio (2002) mostra que a distribuição binomial pertence à família exponencial de distribuições:

$$f(y; m, \pi) = \exp\left\{ln\left[\binom{m}{y}\pi^{y}(1-\pi)^{m-y}\right]\right\}$$

$$f(y; m, \pi) = \exp\left\{ln\binom{m}{y} + y\ ln\,\pi + (m-y)\ ln(1-\pi)\right\}$$

$$= \exp\left\{y\ ln\frac{\pi}{1-\pi} + m\ ln(1-\pi) + ln\binom{m}{y}\right\}.$$

Comparando com a função de distribuição da família exponencial dada em (1.4), $f(y_i; \theta_i, \phi) = \exp\left\{\frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\}$, obtém-se:

$$\begin{aligned} a(\phi) &= 1 \\ \theta &= ln\frac{\pi}{1-\pi} \Longrightarrow \pi = \frac{e^{\theta}}{1+e^{\theta}} \\ b(\theta) &= -\,m\ ln(1-\pi) = m\ ln(1+e^{\theta}) \end{aligned} \qquad (2.2)$$

$$c(y;\phi) = ln\binom{m}{y}.$$

Logo, a distribuição binomial é uma distribuição da família exponencial.

(ii) ***Componente sistemático***

Sendo θ o preditor canônico, isto é, $\eta(\theta) = \theta$, das expressões (1.5) e (2.2) deduz-se que o componente sistemático fica definido por:

$$\eta_i(\theta_i) = \theta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{2.3}$$

(iii) ***Função de ligação canônica***

A função que liga o preditor linear à $\mathrm{E}(Y_i) = m_i\pi_i$ é obtida de (2.2) e (2.3):

$$\mathrm{logit}(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

Logo, $g(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right)$ – que é a função de ligação – é monótona e diferenciável $\forall\pi_i \in]0,1[$.

A Tabela 1 apresenta os componentes das distribuições de Bernoulli, binomial e binomial para proporções associados à família exponencial de distribuições.

Tabela 1 –. Distribuições de Bernoulli, binomial e binomial para proporções na família exponencial

	Bernoulli	Binomial	Binomial para proporções
Notação	Bernoulli(π)	B(m, π)	B(m, π)/m
Log-densidade	$y\ ln\frac{\pi}{1-\pi} + ln(1-\pi)$	$y\ ln\frac{\pi}{1-\pi} + m\ ln(1-\pi) + ln\binom{m}{y}$	$m\left[y\ ln\frac{\pi}{1-\pi} + ln(1-\pi)\right] + ln\binom{m}{my}$
Suporte de Y	0, 1	0, 1, …, m	$z/m,\ z \in \{0, 1, \ldots, m\}$
Parâmetro de dispersão, ϕ	-	1	m^{-1}
$b(\theta)$	$ln(1+e^{\theta})$	$m\ ln(1+e^{\theta})$	$ln(1+e^{\theta})$
$c(y; \phi)$	-	$ln\binom{m}{y}$	$ln\binom{m}{my}$
$\mu(\theta)$= E(Y; θ)	$\frac{e^{\theta}}{1+e^{\theta}} = \pi$	$m\frac{e^{\theta}}{1+e^{\theta}} = m\pi$	$\frac{e^{\theta}}{1+e^{\theta}} = \pi$
Ligação canônica	$ln\left(\frac{\mu}{1-\mu}\right) = ln\left(\frac{\pi}{1-\pi}\right)$	$ln\left(\frac{\mu}{m-\mu}\right) = ln\left(\frac{\pi}{1-\pi}\right)$	$ln\left(\frac{\mu}{1-\mu}\right) = ln\left(\frac{\pi}{1-\pi}\right)$
Função de variância, $V(\mu)$	$\mu(1-\mu) = \pi(1-\pi)$	$\frac{\mu}{m}(m-\mu) = m\pi(1-\pi)$	$\mu(1-\mu) = \pi(1-\pi)$

2.4 Características do modelo logístico

Segundo Garson (2010) a regressão logística é uma técnica estatística que tem por objetivo produzir, a partir de um conjunto de observações, um modelo que permita a estimação de valores tomados por uma variável categórica binária, a partir de uma série de variáveis explicativas contínuas ou binárias. Através dela pode-se determinar o percentual de variação na variável dependente que é explicado pelas variáveis independentes, predizer a *chance* de ocorrêcia de sucesso a partir de um conjunto de variáveis explicativas, determinar o efeito das variáveis independentes sobre a dependente, classificar a importância relativa das variáveis independentes, avaliar interações e compreender o impacto das variáveis explicativas (*razão de chances*).

Como a regressão linear ordinária, a regressão logística requer que as observações sejam independentes. Mas, diversamente, não exige linearidade na relação entre as variáveis independentes e a dependente, a qual só é obtida pela função de ligação; não requer variáveis normalmente distribuídas nem homocedasticidade (item 2.2).

Fox (2008) ressalta outras semelhanças entre o modelo de regressão logística e o regressão linear ordinário: a estimativa dos parâmetros regressores é feita através do método da máxima verossimilhança em ambos os casos; o teste de Wald e o teste da razão de verossimilhança (TRV) para os coeficientes comparam-se ao teste *t* e ao teste *F*; a *deviance* do modelo logístico é análoga à soma dos resíduos de quadrados mínimos para o modelo linear ordinário.

Outra característica importante do modelo logístico, lembrada por Agresti (2007) é sua aplicação tanto em estudos prospectivos (dados experimentais) como em estudos retrospectivos (dados observacionais), nos quais a razão de chances é sempre possível de ser obtida e interpretada.

2.5 Interpretação dos parâmetros do modelo logístico

Seja o modelo logístico de regressão

$$\mathrm{E}(Y_i) = \pi_i = \frac{e^{\boldsymbol{x_i^T \beta}}}{1 + e^{\boldsymbol{x_i^T \beta}}}$$

com função de ligação

$$\text{logit}(\pi_i) = ln\frac{\pi_i}{1-\pi_i} = \boldsymbol{x_i^T\beta}, \tag{2.4}$$

no qual $\boldsymbol{x_i^T} = \begin{bmatrix}1 & x_{i1} & x_{i2} & \dots & x_{i,p-1}\end{bmatrix}$ é o vetor que contém as variáveis explicativas e $\boldsymbol{\beta}^T = \begin{bmatrix}\beta_0 & \beta_1 & \dots & \beta_{p-1}\end{bmatrix}$ é o vetor de parâmetros.

Quatro interpretações importantes podem ser feitas neste modelo: o nível mediano de efetividade, os pontos de estabilização, a chance e a razão de chances.

(i) Nível mediano de efetividade

Agresti (2007) apresenta uma interpretação para a inclinação da reta tangente à curva do modelo logístico (Figura 6), para um único parâmetro regressor, com objetivo de estabelecer o nível mediano de efetividade.

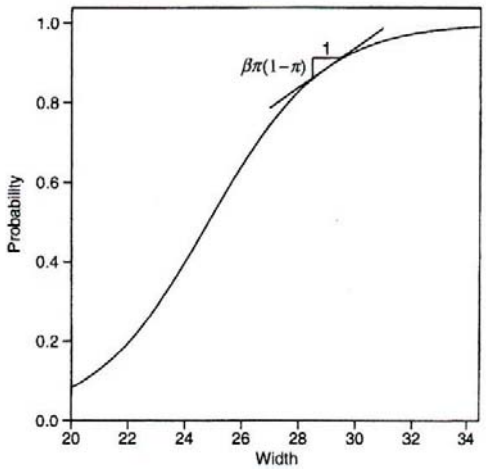


Figura 6 – Aproximação linear para a curva de regressão logística
Fonte: Agresti, 2007, p. 100.

Seja

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \tag{2.5}$$

cuja derivada primeira de $\pi(x)$ em relação a x é dada por:

$$\begin{aligned}\pi'(x) &= \frac{\beta e^{\beta_0+\beta_1 x}}{(1+e^{\beta_0+\beta_1 x})^2} \\ &= \beta \cdot \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \cdot \frac{1}{1+e^{\beta_0+\beta_1 x}}.\end{aligned} \tag{2.6}$$

Mas,

$$1-\pi(x) = 1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{\beta_0+\beta_1 x}}. \tag{2.7}$$

Substituindo (2.5) e (2.7) em (2.6), obtém-se que

$$\pi'(x) = \beta\, \pi(x)[1 - \pi(x)],$$

que é a inclinação da reta tangente à curva para um dado valor de x.

Por consequência, a inclinação da reta tangente para $\pi(x) = 0{,}90$ é igual à de $\pi(x) = 0{,}10$. A maior inclinação ocorre quando $\pi(x) = 0{,}50$ e, neste caso, $x = -\beta_0/\beta_1$, como mostrado a seguir:

$$\begin{aligned}
&\text{logit}(0{,}50) = \beta_0 + \beta_1 x \Leftrightarrow \\
&\Leftrightarrow \frac{0{,}50}{1 - 0{,}50} = e^{\beta_0 + \beta_1 x} \\
&\Leftrightarrow \ln 1 = \ln e^{\beta_0 + \beta_1 x} \\
&\Leftrightarrow 0 = \beta_0 + \beta_1 x \\
&\Leftrightarrow x = -\frac{\beta_0}{\beta_1}.
\end{aligned}$$

Este valor de x é o *nível mediano de efetividade*, EL_{50}, e representa o nível para o qual cada resposta (0 ou 1) tem 50% de probabilidade de ocorrer. Reportando ao ensaio dose-resposta, o nível mediano de efetividade seria o valor da dose para a qual haveria a possibilidade de ser efetiva (sucesso) para 50% dos casos e de não ser efetiva (falha) para os outros 50%. No gráfico da sigmoide, o ponto gerado pelo nível mediano de efetividade é também o ponto de inflexão (ponto A na Figura 7).

(ii) Pontos de estabilização

A curva sigmoide possui características interessantes (PIRES, 2008) representadas pelos pontos A, B e C dados por $(x_{\mathrm{A}}, \pi(x_{\mathrm{A}}))$, $(x_{\mathrm{B}}, \pi(x_{\mathrm{B}}))$ e $(x_{\mathrm{C}}, \pi(x_{\mathrm{C}}))$, respectivamente, como mostrado na Figura 7. No ponto A a curva muda de concavidade (ponto de inflexão) e nos pontos B e C a curva apresenta mudança de direção (ponto de deflexão).

Na sigmoide, o ponto de deflexão está associado ao ponto de curvatura máxima e é também chamado de ponto de estabilização. Em termos estatísticos, os pontos de estabilização dividem a sigmoide em regiões nas quais a relação probabilidade $\pi(x)$ e variável explicativa *x* é diferente: uma alteração fixa para *x* provoca uma variação muito mais acentuada na probabilidade $\pi(x)$ para valores de x entre x_{B} e x_{C} do que para pontos fora desse intervalo.

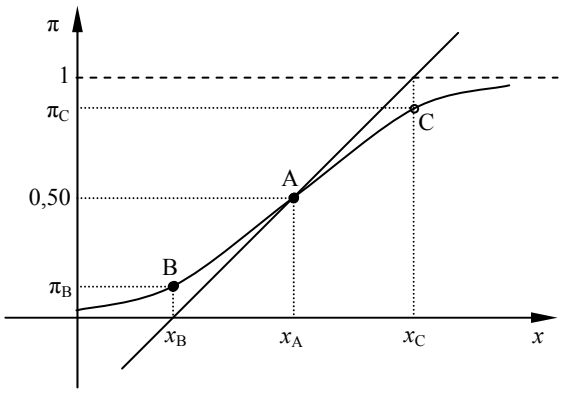


Figura 7 – Esquema gráfico para cálculo do ponto de estabilização
Fonte: baseado em Venegas, Harris e Simon, 1998, p.390.

Os pontos de estabilização de uma sigmoide qualquer podem ser calculados a partir da seguinte sequência de passos (VENEGAs; HARRIS; SIMON, 1998) (Figura 7): (i) determinação da reta tangente que passa pelo ponto de inflexão, (ii) determinação das assíntotas horizontais da função, (iii) determinação dos pontos de interseção entre as assíntotas e a reta tangente (x_B e x_C), e (iv) avaliação dos valores de x_B e x_C na função da sigmoide (π_B e π_C, no gráfico dado).

Como no item precedente, será considerada a curva logística com um único parâmetro regressor β_1, (expressão (2.5)), cujo ponto de inflexão é dado por

$$\mathrm{A} = \left(-\frac{\beta_0}{\beta_1}, \frac{1}{2}\right).$$

Considerando $\beta > 0$ (para $\beta < 0$ é similar), tem-se que o coeficiente angular, *m*, da reta tangente à curva no ponto A é dado por

$$m = \pi'\left(-\frac{\beta_0}{\beta_1}\right) = \frac{\beta e^{\beta_0+\beta_1\left(-\frac{\beta_0}{\beta_1}\right)}}{\left(1+e^{\beta_0+\beta_1\left(-\frac{\beta_0}{\beta_1}\right)}\right)^2} = \frac{\beta_1}{4},$$

sendo, então a equação da reta tangente

$$y = \frac{\beta_1 x}{4} + \frac{\beta_0 + 2}{4}.$$

No caso da função logística, as assíntotas horizontais são dadas por $y = 0$ e $y = 1$, cujas interseções com a reta tangente podem ser calculadas pela resolução dos sistemas

$$\begin{cases} y = 0 \\ y = \dfrac{\beta_1 x}{4} + \dfrac{\beta_0 + 2}{4} \end{cases} \quad \text{e} \quad \begin{cases} y = 1 \\ y = \dfrac{\beta_1 x}{4} + \dfrac{\beta_0 + 2}{4} \end{cases},$$

gerando soluções, respectivamente iguais a

$$x_{\mathrm{B}} = \frac{-\beta_0 - 2}{\beta_1} \quad \text{e} \quad x_{\mathrm{C}} = \frac{-\beta_0 + 2}{\beta_1}.$$

As probabilidades para esses valores de *x* são dadas por:

$$\pi\left(\frac{-\beta_0 - 2}{\beta_1}\right) = \frac{e^{-2}}{1 + e^{-2}} \quad \text{e} \quad \pi\left(\frac{-\beta_0 + 2}{\beta_1}\right) = \frac{e^{2}}{1 + e^{2}}.$$

Logo,

$$\pi\left(\frac{-\beta_0 - 2}{\beta_1}\right) \approx 0{,}12 \quad \text{e} \quad \pi\left(\frac{-\beta_0 + 2}{\beta_1}\right) \approx 0{,}88.$$

Assim,

$$\mathrm{B} = \left(\frac{-\beta_0 - 2}{\beta_1},\ 0{,}12\right) \quad \text{e} \quad \mathrm{C} = \left(\frac{-\beta_0 + 2}{\beta_1},\ 0{,}88\right),$$

que são os pontos de estabilização.

Observa-se que as probabilidades acima independem dos valores de β_0 e β_1, o que quer dizer que em todas as curvas logísticas com um único parâmetro regressor os valores das probabilidades nos pontos de estabilização são sempre os mesmos. Devido à simetria da curva em relação ao nível mediano de efetividade, também se observa que

$$\pi\left(\frac{-\beta_0 + 2}{\beta_1}\right) + \pi\left(\frac{-\beta_0 - 2}{\beta_1}\right) = 1.$$

(iii) Chance (odds)

Chama-se de chance ou risco de ocorrência do evento $Y_i = 1$ dado $\boldsymbol{X} = \boldsymbol{x_i}$ à razão entre a probabilidade de ocorrência e a de não ocorrência do evento:

$$\text{chance}(Y_i = 1 | \boldsymbol{X} = \boldsymbol{x_i}) = \frac{\pi_i}{1 - \pi_i}. \tag{2.8}$$

Comparando esta expressão com a expressão (2.4), verifica-se que a função de ligação logito nada mais é que o logaritmo neperiano da chance de ocorrência de um evento. Assim, usando as expressões (2.2) e (2.3) e substituindo em (2.8) deduz-se que em regressão logística a chance de ocorrência do evento $Y_i = 1$ dado $\boldsymbol{X} = \boldsymbol{x_i}$ é dada por:

$$\text{chance}(Y_i = 1 | \boldsymbol{X} = \boldsymbol{x_i}) = \frac{\pi_i}{1 - \pi_i} = \frac{\dfrac{e^{\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}}}{\dfrac{1}{1 + e^{\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}}}$$

$$\text{chance}(Y_i = 1 | \boldsymbol{X} = \boldsymbol{x_i}) = e^{\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}}.$$

Fazendo $\boldsymbol{\beta} = [\beta_0 \quad \cdots \quad \beta_{p-1}]^{\mathrm{T}}$ e se $\beta_r = 0$ para $r = 1, \ldots, p-1$, então as

chances serão iguais para todo $\boldsymbol{X} = \boldsymbol{x_i}$, isto é, $\text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i}) = e^{\beta_0},\ \forall \boldsymbol{x_i}$.

Agresti (2007) ressalta que chance não é probabilidade, mas uma razão entre probabilidades, podendo, assim, assumir qualquer valor não negativo. Quanto mais a probabilidade de ocorrência de um evento se aproxima de zero, mais a sua chance de ocorrência se aproxima de zero também; por outro lado, quanto mais a probabilidade de ocorrência de um evento se aproxima de 1, maior se torna a sua chance de ocorrência, teoricamente tendendo ao infinito.

Quando a chance de ocorrência de um evento é igual a 1 isto quer dizer que a probabilidade de ocorrência ou não de um evento são iguais, isto é, ambas iguais a 0,50, o que representa um equilíbrio entre sucesso e fracasso. Assim o valor 1 para a chance pode ser tomado como base de análise. Se a chance for menor do que 1, então a probabilidade de sucesso é menor que a probabilidade de fracasso; se a chance for maior do que 1, então a probabilidade de sucesso é maior que a probabilidade de fracasso.

Para melhor entender o conceito de chance, considere-se o exemplo descrito a seguir.

Na calçada de uma rua há um buraco que tem provocado alguns acidentes. Um morador local verificou que 80% dos cegos e 10% dos não cegos caem no buraco. Pretende-se determinar a chance que têm um cego e um não cego de caírem no buraco.

Considerando como "sucesso" cair no buraco e os dados do problema, obtém-se:

$$\text{P}(Y = 1|X = \text{cego}) = 0{,}80 \text{ e } \text{P}(Y = 0|X = \text{cego}) = 1 - 0{,}80 = 0{,}20$$

$$\text{P}(Y = 1|X = \text{não cego}) = 0{,}10 \text{ e } \text{P}(Y = 0|X = \text{não cego}) = 1 - 0{,}10 = 0{,}90$$

Portanto,

$$\text{chance}(Y = 1|X = \text{cego}) = \frac{0{,}80}{0{,}20} = \frac{8}{2} = \frac{4}{1}$$

$$\text{chance}(Y = 1|X = \text{não cego}) = \frac{0{,}10}{0{,}90} = \frac{1}{9} \approx 0{,}11$$

Interpretando as chances obtidas, verifica-se que um cego em particular tem chance de 4 para 1 de cair no buraco comparado a não cair, isto é, em 10 vezes que passar por aquela calçada é de se esperar que 8 vezes o cego caia no buraco contra 2 para não cair. Para um não cego particular, de 10 vezes que passar por este mesmo lugar, a chance de cair é de 1 contra 9 de não cair. Como era previsível, a chance de um cego cair no buraco é bem maior que a chance de um não cego.

(iv) Razão de chances (odds ratio)

A razão de chances é utilizada para comparar as chances de dois indivíduos, avaliando o quanto a chance de um indivíduo é maior ou menor que a chance de outro. Por definição, a razão de chances, *OR*, entre dois indivíduos quaisquer x_i e x_j é dada por:

$$OR(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i})}{\text{chance}(Y_j = 1|\boldsymbol{X} = \boldsymbol{x_j})}.$$

Supondo que $OR(\boldsymbol{x_i}, \boldsymbol{x_j}) = k,\ k \in \mathbb{R}$, então:

$$\frac{\text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i})}{\text{chance}(Y_j = 1|\boldsymbol{X} = \boldsymbol{x_j})} = k \Longrightarrow \text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i}) = k \cdot \text{chance}(Y_j = 1|\boldsymbol{X} = \boldsymbol{x_j}).$$

Isto quer dizer que a chance de ocorrência do evento $Y_i = 1$ dado $\boldsymbol{X} = \boldsymbol{x_i}$ é k vezes a chance de ocorrência do evento $Y_j = 1$ dado $\boldsymbol{X} = \boldsymbol{x_j}$.

Fazendo uso do exemplo anterior, tem-se que a razão de chances de cair no buraco entre um cego e um não cego é dada por:

$$OR(\text{cego, não cego}) = \frac{\text{chance}(Y = 1|X = \text{cego})}{\text{chance}(Y = 1|X = \text{não cego})}$$

$$OR(\text{cego, não cego}) = \frac{4/1}{1/9}$$

$$OR(\text{cego, não cego}) = \frac{36}{1}$$

Logo, a chance de um cego cair no buraco é 36 vezes maior do que a de um não cego.

Na regressão logística, a razão de chances leva à interpretação das componentes do vetor $\boldsymbol{\beta}$. Supondo que os vetores $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$ diferem entre si de uma unidade apenas na componente s, isto é, a componente s do vetor $\boldsymbol{x_j}$ é igual a $(x_{is} + 1)$ e as demais são todas iguais: $\boldsymbol{x_i^T} = [1 \ \ x_{i1} \ \ \ldots \ \ x_{is} \ \ldots \ \ x_{i,p-1}]$ e $\boldsymbol{x_j^T} = [1 \ \ x_{i1} \ \ \ldots \ \ x_{is} + 1 \ \ldots \ \ x_{i,p-1}]$. Define-se a razão de chances, *OR*, entre dois vetores $\boldsymbol{x_j}$ e $\boldsymbol{x_i}$ por:

$$\begin{aligned} OR(\boldsymbol{x_j}, \boldsymbol{x_i}) &= \frac{\text{chance}(Y_j = 1|\boldsymbol{X} = \boldsymbol{x_j})}{\text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i})} \\ &= \frac{\exp[\beta_0 + \beta_1 x_1 + \cdots + \beta_s(x_{is} + 1) + \cdots \beta_{p-1} x_{i,p-1}]}{\exp[\beta_0 + \beta_1 x_1 + \cdots + \beta_s x_{is} + \cdots \beta_{p-1} x_{i,p-1}]} \\ &= \exp\{[\beta_0 + \beta_1 x_1 + \cdots + \beta_s x_{is} + \beta_s + \cdots \beta_{p-1} x_{i,p-1}] - \\ &\quad -[\beta_0 + \beta_1 x_1 + \cdots + \beta_s x_{is} + \cdots \beta_{p-1} x_{i,p-1}]\} \end{aligned}$$

Logo,

$$OR(\boldsymbol{x_j}, \boldsymbol{x_i}) = e^{\beta_s}, \tag{2.9}$$

isto é, $\text{chance}(Y_j = 1|\boldsymbol{X} = \boldsymbol{x_j}) = e^{\beta_s} \cdot \text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i})$, ficando bem clara a interpretação do parâmetro β_s: a chance de sucesso dado $\boldsymbol{X} = \boldsymbol{x_j}$ é e^{β_s} vezes a chance de sucesso dado $\boldsymbol{X} = \boldsymbol{x_i}$. Ainda que a componente s do vetor $\boldsymbol{x_i}$ seja binária ou indicadora, a interpretação do parâmetro β_s continua sendo possível e de fácil compreensão, como pode ser visto no exemplo a seguir, usado por Ogliari ([200-]).

Um estudo na área da saúde está investigando um surto epidêmico de uma doença transmitida por um mosquito. Indivíduos foram aleatoriamente selecionados em dois setores de uma cidade para determinar se a pessoa tinha recentemente contraído a doença em estudo. Três variáveis preditoras foram incluídas no estudo: idade, status socioeconômico da família e o setor da cidade. A idade (X_1) é uma variável quantitativa; o status socioeconômico é uma variável com 3 categorias, sendo representada pelo par de variáveis indicadoras (X_2, X_3), onde (0, 0) representa a classe alta (tomada como referência, por se esperar menor taxa de casos da doença nesta classe), (1, 0) a classe média e (0, 1) a classe baixa; para o setor da cidade também foi usada uma variável indicadora X_4, onde 0 representa o setor 1 (referência, por ter apresentado menos casos da doença) e 1 para o setor 2. A variável resposta *Y* foi codificada como 1 se a doença estava presente, e 0 em caso contrário. O primeiro propósito da análise foi verificar a força de associação entre as variáveis preditoras e a probabilidade de uma pessoa ter contraído a doença.

Foi ajustado o seguinte modelo logístico:

$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}$$

sendo $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

Foram obtidas as seguintes estimativas de máxima verossimilhança da função de regressão logística: $\hat{\beta}_0$ = -2,3129, $\hat{\beta}_1$ = 0,02975, $\hat{\beta}_2$ = 0,4088, $\hat{\beta}_3$ = -0,30525, $\hat{\beta}_4$ = 1,5747.

Para fazer a interpretação dos parâmetros é preciso que apenas uma das variáveis sofra modificação, mantendo-se as demais fixas.

A razão de chances entre dois indivíduos com diferença de 1 ano na idade, mas de mesmo status socioeconômico e morador do mesmo setor da cidade, é dada por:

$$\widehat{OR} = e^{\hat{\beta}_1} = e^{0,02975} = 1,03.$$

Este resultado mostra que cada ano que se adiciona à idade aumenta em 3% a chance de uma pessoa ter contraído a doença, mantendo-se fixos o status socioeconômico e o setor da moradia.

A razão de chances entre indivíduos moradores do setor 2 e do setor 1, mas com mesma idade e status socioeconômico, é dada por:

$$\widehat{OR} = e^{\widehat{\beta}_4} = e^{1,5747} = 4,83.$$

O que quer dizer que a chance de um morador do setor 2 ter contraído a doença é quase 5 vezes maior do que um morador do setor 1.

A razão de chances entre um indivíduo da classe média e outro da classe alta, mas ambos da mesma idade e morando no mesmo setor da cidade, é dada por:

$$\widehat{OR} = e^{\widehat{\beta}_2} = e^{0,4088} = 1,51.$$

Este valor indica que a chance de uma pessoa da classe média ter contraído a doença é 50% maior que uma pessoa da classe alta.

A razão de chances entre um indivíduo da classe baixa e outro da classe alta, mas ambos da mesma idade e morando no mesmo setor da cidade, é dada por:

$$\widehat{OR} = e^{\widehat{\beta}_3} = e^{-0,30525} = 0,74,$$

indicando que a chance de uma pessoa da classe baixa ter contraído a doença é 26% menor que de uma pessoa da classe alta. É importante observar que neste percentual a classe alta foi tomada como base. Fazendo o contrário, ter-se-ia 36% a mais chance para um indivíduo da classe alta em relação à classe baixa. Este valor teria sido diretamente obtido se a razão de chances tivesse sido calculada do indivíduo da classe alta para o da classe baixa:

$$\widehat{OR} = 1/e^{\widehat{\beta}_3} = e^{0,30525} = 1,36.$$

A observação acima representa uma das propriedades da razão de chances, que será vista mais adiante.

Seria ainda possível calcular a razão de chances entre dois indivíduos tomados da classe baixa e da classe média, moradores do mesmo setor e com a mesma idade. Usando-se a definição de razão de chances obtém-se a expressão:

$$\widehat{OR} = e^{\widehat{\beta}_2 - \widehat{\beta}_3} = e^{0,7145} = 2,04,$$

a qual mostra que a chance de um indivíduo da classe média ter contraído a doença é o dobro (100% maior) da chance de um indivíduo da classe baixa.

Propriedades da razão de chances (AGRESTI, 2007):

a) A razão de chances pode assumir qualquer valor não negativo.

b) Se $OR(\boldsymbol{x_j}, \boldsymbol{x_i}) = 1$, então $\text{chance}(Y_j = 1 | \boldsymbol{X} = \boldsymbol{x_j}) = \text{chance}(Y_i = 1 | \boldsymbol{X} = \boldsymbol{x_i})$ e, consequentemente, $\pi_j = \pi_i$. Este caso representa a situação de independência entre $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$.

c) Se $OR(\boldsymbol{x_j}, \boldsymbol{x_i}) > 1$, então $\pi_j > \pi_i$.

d) Se $OR(\boldsymbol{x_j}, \boldsymbol{x_i}) < 1$, então $\pi_j < \pi_i$.

e) As razões de chances $OR(\boldsymbol{x_j}, \boldsymbol{x_i})$ e $OR(\boldsymbol{x_i}, \boldsymbol{x_j})$ representam a mesma força de associação, porém em direções opostas, sendo $OR(\boldsymbol{x_i}, \boldsymbol{x_j}) = 1/OR(\boldsymbol{x_j}, \boldsymbol{x_i})$.

2.6 Estimação dos parâmetros

Considerando o modelo logístico com componente aleatório $Y_i \sim \text{Binomial}(m_i, \pi_i)$, e usando os resultados obtidos nos itens 1.4.2, 1.5 e 2.3, tem-se que:

$$\text{logit}(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i}$$

$$\text{E}(Y_i) = m_i \pi_i = \mu_i$$

$$\theta_i = \ln \frac{\pi_i}{1 - \pi_i} \Longrightarrow \theta_i = \ln \frac{\mu_i}{m_i - \mu_i}$$

$$a(\phi) = \phi = 1$$

$$b(\theta_i) = m_i \ln\left(1 + e^{\theta_i}\right)$$

$$b'(\theta_i) = m_i \frac{e^{\theta_i}}{1 + e^{\theta_i}} = \mu_i$$

$$b''(\theta_i) = m_i \frac{e^{\theta_i}}{(1 + e^{\theta_i})^2} = V_i \Longrightarrow V_i(\mu_i) = \frac{\mu_i}{m_i}(m_i - \mu_i)$$

$$\eta_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$$

$$g(\mu_i) = \eta_i = \ln \frac{\mu_i}{m_i - \mu_i} = \theta_i \quad \text{(função de ligação canônica)}$$

$$g^{-1}(\eta_i) = \mu_i = \frac{m_i e^{\eta_i}}{1 + e^{\eta_i}}$$

$$g'(\mu_i) = \frac{m_i}{\mu_i (m_i - \mu_i)} = \frac{1}{V_i}$$

Os resultados apresentados acima são necessários para a construção das matrizes envolvidas na estimação do vetor de parâmetros $\boldsymbol{\beta}$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix},$$

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p \\ x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p \\ \vdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{\beta} \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\beta} \end{bmatrix}$$

As seguintes matrizes são necessárias para o desenvolvimento do algoritmo de estimação:

$\boldsymbol{\mu}_{n\times 1}$, onde $\mu_i = m_i \dfrac{e^{\eta_i}}{1+e^{\eta_i}}$

$\boldsymbol{V}_{n\times 1}$, onde $V_i = \dfrac{\mu_i(m_i-\mu_i)}{m_i}$

$\mathbf{W}_{n\times n}$, $\mathbf{W} = \mathrm{diag}(V_1, V_2, \ldots, V_n)$

$\mathbf{G}_{n\times n}$, $\mathbf{G} = \mathbf{W}^{-1} = \mathrm{diag}(1/V_1, 1/V_2, \ldots, 1/V_n)$

(i) ***Estimativa inicial para* z e $\boldsymbol{\beta}$**

$$\mu_i = y_i$$
$$\mathrm{z}_i^{(1)} = \eta_i^{(1)} = g(y_i) = \ln\frac{y_i}{m_i - y_i} \qquad (2.9)$$
$$\boldsymbol{\beta}^{(2)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(1)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(1)}\boldsymbol{\eta}^{(1)}, \text{ sendo } \; w_i{}^{(1)} = \frac{y_i(m_i - y_i)}{m_i}$$

(ii) ***Cálculo de* $\boldsymbol{\beta}^{(m+1)}$**

(1º) Obtenção das estimativas de $\boldsymbol{\eta}$ e $\boldsymbol{\mu}$

$$\eta_i{}^{(m)} = \sum_{r=1}^{p} x_{ir}\beta_r{}^{(m)}$$
$$\mu_i{}^{(m)} = \frac{m_i e^{\eta_i{}^{(m)}}}{1+e^{\eta_i{}^{(m)}}}$$

(2º) Obtenção das estimativas de **z** e **W**

$$z_i^{(m)} = \eta_i^{(m)} + \frac{m_i}{\mu_i{}^{(m)}(m_i - \mu_i{}^{(m)})}\left(y_i - \mu_i^{(m)}\right)$$

$$= \eta_i^{(m)} + \frac{\mu_i{}^{(m)}}{1 + e^{\eta_i{}^{(m)}}}\left(y - \mu_i^{(m)}\right)$$

$$w_i{}^{(m)} = \frac{\mu_i{}^{(m)}(m_i - \mu_i{}^{(m)})}{m_i}$$

$$= \frac{\mu_i{}^{(m)}}{1 + e^{\eta_i{}^{(m)}}}$$

(3º) Obtenção da estimativa de $\boldsymbol{\beta}$

$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{z}^{(m)},\ m \geq 2$, até que

$$\sum_{r=1}^{p}\left(\frac{\beta_r^{(m+1)} - \beta_r^{(m)}}{\beta_r^{(m)}}\right)^2 < \xi$$

Como pode ser observado em (2.9), se para algum i, $y_i = 0$ ou $y_i = m_i$, não será possível o cálculo do valor inicial de $\mathbf{z}$. Esse problema é resolvido através de um ajuste para o valor de $\boldsymbol{y}$ chamado de penalização.

2.7 Outras estimativas

A partir de $\widehat{\boldsymbol{\beta}}$, a estimativa do vetor de parâmetros $\boldsymbol{\beta}$ obtida pelo MNR, pode-se calcular todas as outras estimativas e realizar teste de hipóteses:

a) Vetor preditor linear:

$$\hat{\eta}_i = \sum_{r=1}^{p} x_{ir}\hat{\beta}_r$$

b) Vetor média

$$\hat{\mu}_i = \frac{m_i e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}}$$

c) Matriz de pesos

$$\widehat{w}_i = \frac{\hat{\mu}_i}{1 + e^{\hat{\eta}_i}}$$

d) Vetor escore

$$\widehat{\mathrm{U}}_r = \sum_{i=1}^{n} x_{ir}(y_i - \hat{\mu}_i)$$

e) Matriz de informação de Fisher

$$\widehat{\mathrm{K}}_{rs} = \sum_{i=1}^{n} x_{ir}\widehat{w}_i x_{is}$$

f) Chance

$$\mathrm{logit}(\hat{\pi}_i) = ln\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}} = \hat{\eta}_i$$

$$\mathrm{chance}(\hat{\pi}_i) = e^{\hat{\eta}_i}$$

g) Razão de chances

$$\widehat{OR}(\boldsymbol{x_i}, \boldsymbol{x_j}) = e^{(x_i - x_j)\widehat{\boldsymbol{\beta}}},$$

CAPÍTULO 3 – REGRESSÃO LOGÍSTICA COM EFEITO ALEATÓRIO

Dados binários correlacionados ocorrem frequentemente em muitos campos de estudo e podem ter diferentes delineamentos, como os dados longitudinais onde as respostas são mensuradas no mesmo indivíduo repetidamente ao longo do tempo, estudos genéticos onde membros de uma família estão intimamente relacionados, ou estudos em agricultura onde os agrupamentos são naturalmente definidos. Não importando como o estudo foi desenvolvido, é preciso incorporar a correlação no modelo para obter as estimativas corretas e tirar as conclusões apropriadas (LARSEN et al, 2000). Tal incorporação dá origem aos modelos mistos que apresentam na sua estrutura tanto parâmetros para os efeitos fixos como para os efeitos aleatórios.

Pinheiro e Bates (2000) afirmam que modelos mistos, proporcionam uma ferramenta poderosa para análise de dados em grupo, cuja popularidade crescente é explicada pela flexibilidade na modelagem da correlação intragrupo.

Porém, há muitas controvérsias sobre o conceito de efeito aleatório e, consequentemente sobre seu uso, modelagem, análise e inferência. Em alguns casos, como mostra Merlo (2003), o efeito aleatório é encarado apenas como um fator de perturbação que pode ser ajustado na análise e não explicitamente investigado. Outros problemas aparecem na modelagem multinível, onde, segundo Larsen e Merlo (2005) a partição da variância em diferentes níveis é condição *sine qua non*, mas, contrariamente ao que acontece com as variáveis com distribuição normal, as componentes de variância para as variáveis respostas dicotômicas são difíceis de investigar. Forçando a interpretação pelos esquemas clássicos pode-se levar à impropriedade e à perda da informação.

Em regressão logística, Larsen representa um expoente no estudo do efeito aleatório, fazendo considerações, propondo e difundindo um método prático de interpretação deste tipo de modelo: a razão de chances mediana e seu intervalo percentílico (LARSEN et al., 2000; LARSEN; MERLO, 2005; LARSEN, 2006; MERLO et al., 2006). Seu objetivo é apresentar uma interpretação para modelos logísticos mistos com a mesma intuição e facilidade que os modelos de efeito fixo apresentam através da razão de chances. Detalhes desta proposta serão vistas no desenvolvimento deste capítulo.

Larsen et al. (2000) apresentam três possibilidades diferentes de modelagem conjunta: modelos *threshold* (ASHFORD e SOWDEN, 1972), método GEE (ZEGER et

al., 1988; LINDSEY e LAMBERT, 1988) e modelos de efeito aleatório (WILLIAMS, 1975; STIRATELLI et al., 1984). Como em seus estudos, neste também será enfocado o terceiro caso, através de um modelo de regressão logística com efeito aleatório.

3.1 Efeito aleatório

Modelos mistos de regressão logística são aqueles que apresentam na sua estrutura tanto efeitos fixos como efeitos aleatórios. Os efeitos fixos são parâmetros associados a uma população como um todo ou com certo nível de fator experimental. Os efeitos aleatórios estão associados com uma unidade experimental individual tomada aleatoriamente da população. Segundo Pinheiro e Bates (2000) modelos mistos são usados primariamente para descrever relações entre a variável resposta e alguma explicativa dos dados, os quais podem ser agrupados segundo um ou mais fatores de classificação. Constituem exemplos de dados agrupados os dados longitudinais, medidas repetidas sobre os dados, dados multinível e delineamentos em bloco.

Pinheiro e Bates (2000) afirmam que determinar o que é efeito fixo ou efeito aleatório num experimento pode depender do objetivo do mesmo. Este pensamento coaduna com o de Littell, Freund e Spector (1991) que afirmam que uma diferença chave entre efeito aleatório e fixo é o tipo de informação que se quer analisar dos efeitos. No caso de efeitos fixos, normalmente deseja-se fazer comparações entre seus níveis. No caso de efeitos aleatórios, deseja-se saber o quanto este influencia a variância da variável dependente.

Para elucidar tal pensamento, Pinheiro e Bates (2000) fazem uso de um exemplo envolvendo trilhos de uma ferrovia, cujas considerações estão resumidas nos próximos parágrafos.

Seis trilhos são tomados aleatoriamente e testados três vezes cada para medir o tempo que certo tipo de onda ultrassônica leva para percorrer toda a extensão do trilho. As quantidades que os engenheiros podem estar interessados em estimar neste experimento são o tempo médio de travessia num trilho "típico", a variação na média de travessia entre os trilhos, e a variação observada na travessia para um único trilho. Se o que se deseja é fazer inferências sobre aqueles particulares níveis de fatores que foram usados no experimento, o experimento pode ser analisado por um modelo de efeito fixo;

caso o interesse resida em fazer inferências sobre a população da qual se extraiu esses níveis, então um modelo misto deve ser utilizado.

Ignorando a estrutura de dados em grupo, tem-se o seguinte modelo de efeito fixo:

$$y_{ij} = \beta + \epsilon_{ij}, \quad i = 1, \dots, M, \;\; j = 1, \dots, n_i,$$

em que y_{ij} é o tempo de travessia observado na j-ésima observação do trilho *i,* β é o tempo médio de travessia; ϵ_{ij} é o erro associado à observação ij, iid, com distribuição N(0, σ^2); M é o número de trilhos – igual a 6, neste exemplo – e n_i é o número de observações sobre o trilho i – igual a 3 para todos os trilhos.

Neste modelo o "efeito de grupo" está incorporado ao resíduo, com idêntico significado para cada trilho, levando a uma estimativa exagerada da variabilidade dentro do trilho.

Ainda considerando um modelo de efeito fixo, mas incorporando o "efeito de trilho" para o tempo de travessia apresentando uma média separada para cada trilho:

$$y_{ij} = \beta_i + \epsilon_{ij}, \quad i = 1, \dots, M, \;\; j = 1, \dots, n_i,$$

onde β_i representa a média do tempo de travessia no trilho i. Os erros ϵ_{ij} são assumidos iid, com distribuição N(0, σ^2).

Neste caso, os erros ficam consideravelmente menores que no modelo anterior. Porém, neste modelo só é possível avaliar o efeito de cada trilho, visto que não há uma representação para o conjunto de trilhos.

Se o interesse reside na população de trilhos da qual foi retirada a amostra, um modelo com efeito aleatório contorna este problema, tratando o efeito de trilho como uma variação aleatória em torno da média populacional. O modelo a seguir ajuda a motivar o modelo de efeito aleatório para os dados dos trilhos:

$$y_{ij} = \bar{\beta} + \left(\beta_i - \bar{\beta}\right) + \epsilon_{ij},$$

onde $\bar{\beta} = \sum_{i=1}^{6} \beta_i / 6$ representa o tempo médio de travessia pelos trilhos no experimento.

No modelo de efeito aleatório a média $\bar{\beta}$ é substituída pela média sobre toda a população de trilhos e o desvio $\left(\beta_i - \bar{\beta}\right)$ é substituído por variáveis aleatórias cuja distribuição deverá ser estimada:

$$y_{ij} = \beta + b_i + \epsilon_{ij},$$

sendo β a média de travessia na população de trilhos, b_i é a variável aleatória representando o desvio da média de tempo do i-ésimo trilho em relação à média populacional e ϵ_{ij} representa o desvio do tempo de travessia da observação j no trilho i em relação em relação ao tempo médio do trilho i.

Kreft e De Leeuw (1998) apresentam uma interpretação geométrica para o modelo com efeito aleatório $y_{it} = a_i + bt$ em estudos de crescimento, no qual a_i representa o efeito aleatório entre os indivíduos e b é o efeito fixo do tempo. Assim, o modelo que apresenta intercepto a_i flutuante e inclinação fixa b corresponde a linhas paralelas para diferentes indivíduos *i* (Figura 1(*a*)).

Uma generalização desta interpretação pode ser feita para outros modelos mistos com efeito aleatório no intercepto:

$$y_{ij} = (\beta_0 + \delta_{ij}) + \beta_1 x_i + e_{ij},$$

onde δ_j representa o efeito aleatório dentro do nível *j*.

Neste caso, o intercepto é dado por $(\beta_0 + \delta_j)$ e a inclinação fixa da reta é dada por β_1 (Figura 1(*b*)).

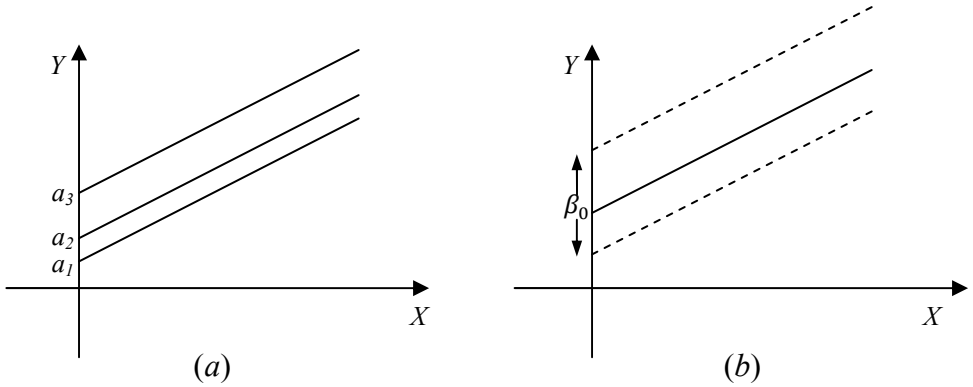


Figura 1 – Efeito aleatório no intercepto: (*a*) $y_{it} = a_i + bt$; (*b*) $y_{\mathrm{ij}} = (\beta_0 + \delta_{\mathrm{j}}) + \beta_1 X_{\mathrm{i}}$

Snijders (2005) afirma que normalmente – embora não necessariamente – um efeito modelado como aleatório é assumido como tendo distribuição normal. Porém, nem sempre isto está de acordo com a realidade, levando a resultados viesados. A alternativa de usar distribuições não normais pode ser complicada, mas já existem métodos de estimação para tal, como os modelos de fragilidade gama. Finalizando, Weber (2011) lembra que toda heterogeneidade não modelada é incorporada ao erro aleatório, inflacionando-o.

3.2 Modelo logístico com efeito aleatório no intercepto

Neste estudo será usado o modelo logístico com efeito aleatório no intercepto. A notação utilizada é a constante em Larsen et al. (2000).

Seja Y_i uma variável aleatória com distribuição de Bernoulli, para $i = 1, ..., N$, probabilidade dada por:

$$P(Y_i = 1|\eta_i) = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}}, \quad \text{sendo } \eta_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\mathbf{u}, \tag{3.1}$$

onde :

$\boldsymbol{\beta}_{p\times 1}$ é o vetor de parâmetros dos efeitos fixos,

$\mathbf{X}_{N\times p}$ é a matriz para os efeitos fixos e $\boldsymbol{x}_i$ é a i-ésima linha dessa matriz,

$\mathbf{u}_{r\times 1}$ é o vetor que contém os efeitos aleatórios, os quais seguem distribuição normal com média $\mathbf{0}$ e matriz de variância $\boldsymbol{\Sigma}$, e

$\mathbf{Z}_{N\times r}$ é a matriz para os efeitos aleatórios e $\boldsymbol{z}_i$ é a i-ésima linha dessa matriz.

O vetor $\mathbf{u}$ pode ser pensado como um vetor de covariáveis não mensuráveis usado como meio de modelar a heterogeneidade ou modelar dados correlacionados.

3.3 Razão de chances

Em regressão logística com efeito aleatório, a razão de chances pode ter a mesma interpretação intuitiva que na regressão logística simples. Porém, o efeito aleatório não é observável, o que faz com que o estimador da razão de chances seja dado por uma fórmula mais complexa.

Larsen et al. (2000), consideram que em presença de efeito aleatório a razão de chances é uma variável aleatória e propõem uma interpretação baseada na mediana, isto é, a razão de chances mediana (MOR), tanto para o efeito aleatório como para o efeito fixo, associando a este último o intervalo percentílico da razão de chances (IOR).

Todos os cálculos e interpretações dados a seguir estão baseados na proposta de Larsen et al. (2000) com acréscimo de detalhes de Larsen e Merlo (2005).

Por definição, a razão de chances entre duas medidas i_1 e i_2 tem a forma seguinte:

$$\text{chance}\left(Y_{i_1} = 1 \middle| \boldsymbol{X} = \boldsymbol{x}_{i_1}\right) = e^{\boldsymbol{\eta}_{i_1}}$$

$$\text{chance}\left(Y_{i_2} = 1 \middle| \boldsymbol{X} = \boldsymbol{x}_{i_2}\right) = e^{\boldsymbol{\eta}_{i_2}}.$$

$$OR = \exp\left(\boldsymbol{\eta}_{i_1} - \boldsymbol{\eta}_{i_2}\right)$$

$$= \exp\left[\left(\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2}\right)\boldsymbol{\beta} + \left(\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2}\right)\mathbf{u}\right]$$

Assim, a OR pode ser escrita como:

$$OR = \exp(\kappa + \omega \cdot v) \tag{3.1}$$

onde $\kappa = (\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2})\boldsymbol{\beta}$, $\omega^2 = (\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2})\boldsymbol{\Sigma}(\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2})^{\mathrm{T}}$ e $v \sim N(0,1)$.

Desta forma, a razão de chances não é mais um parâmetro fixo, mas uma variável aleatória, produzindo reflexos na interpretação do modelo. Possui, também, duas partes distintas: uma associada aos parâmetros de efeito fixo e outra associada ao efeito aleatório.

Para ilustração e esclarecimento dos conceitos será adotado o exemplo das pocilgas, presente em Larsen et al. (2000).

Para investigar a ocorrência de *Ascaris suum* (lombriga) em porcos, amostras de fezes foram retiradas de 1016 porcos de abate e fêmeas na Dinamarca. Dois tipos diferentes de pocilgas foram incluídas no estudo, a convencional e a chamada de SPF. Os dados parciais obtidos a partir de 72 pocilgas convencionais e 36 SPF encontram-se em Larsen et al. (2000). O objetivo foi investigar se os porcos em pocilgas SPF têm um menor risco de serem infectados do que os porcos em pocilgas convencionais e, em caso afirmativo, quantificar esta diferença. Supôs-se que cada pocilga tem seu próprio nível de infecção. Para modelar esta heterogeneidade entre as pocilgas, foi utilizada a regressão logística com efeito aleatório como em (3.1), onde **X** é uma matriz de dimensão 1016×2, com elemento $x_{ik} = 1$, se o i-ésimo porco pertence à pocilga tipo j; **Z** é uma matriz 1016×108, com elemento $z_{ij} = 1$, se se o i-ésimo porco pertence à k-ésima pocilga; e Σ é uma matriz diagonal 108×108, com elementos σ^2. Os parâmetros foram estimados usando o SAS, macro GLIMMIX, obtendo $\hat{\beta}_{\mathrm{SPF}} = -3{,}03$, $\hat{\beta}_{\mathrm{conv}} = -2{,}14$ e $\hat{\sigma}^2 = 1{,}38$.

A interpretação da razão de chances entre as pocilgas convencionais e SPF sendo avaliada apenas por $e^{\widehat{\beta}_{\mathrm{conv}} - \widehat{\beta}_{\mathrm{SPF}}} = 2{,}44$ – como mostrado em (2.9) – não leva em consideração a heterogeneidade das pocilgas explicitamente na interpretação dos efeitos. Larsen et al. (2000) salientam que o tamanho da variância dentro das próprias pocilgas são releventes porque elas determinam a escala na qual os efeitos fixos podem ser julgados.

3.3.1 Razão de chances envolvendo o efeito aleatório

Ao se considerar dois indivíduos com os mesmos valores de covariáveis para os efeitos fixos, obtém-se $\kappa = 0$ na fórmula (3.1), ficando a razão de chances sendo dada

por

$$OR = \exp(\omega \cdot v)$$

Larsen et al. (2000) apresentam algumas considerações sobre a quantificação desta razão de chances, que sendo uma variável aleatória, a ela pode ser atribuída alguma característica distribucional. Resumidamente, avaliar a OR por sua esperança não é um bom caminho, pois esta não captura a heterogeneidade, isto é, não permite medir a variação da OR; avaliá-la por sua variância também é insatisfatório, visto que sua distribuição é assimétrica. Assim, esses autores propõem a avaliação da OR através da razão de chances mediana entre a unidade de mais alto risco (maior channce) e a de mais baixo risco (menor chance), isto é:

$$MOR_{\text{aleat}} = \text{med}\{\exp(|\omega \cdot v|)\} = \exp\left\{\omega \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\},$$

a qual possui propriedades interessantes como:

$$\text{med}\{\exp(-|\omega \cdot v|)\} = [\text{med}\{\exp(|\omega \cdot v|)\}]^{-1} \text{ e}$$

$$\text{med}\{\exp(|\omega \cdot v|)\} = \exp\{\text{med}(|\omega \cdot v|)\}$$

O estimador para a MOR fica dado por:

$$\widehat{MOR}_{\text{aleat}} = \exp\left\{\sqrt{2\hat{\sigma}^2} \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\},$$

onde $\hat{\sigma}^2$ é a variância estimada para o efeito aleatório.

Demonstração:

Fazendo $MOR_{\text{aleat}} = a$ e $T = \exp(|\omega \cdot v|)$, onde $v \sim N(0,1)$, tem-se que

$$P(T \leq a) = P[\exp(|\omega \cdot v|) \leq a]$$

$$P(T \leq a) = P[|\omega \cdot v| \leq \ln a]$$

$$P(T \leq a) = P[-\ln a \leq \omega \cdot v \leq \ln a]$$

$$P(T \leq a) = P\left[-\frac{\ln a}{\omega} \leq v \leq \frac{\ln a}{\omega}\right].$$

Logo,

$$F_T(a) = \Phi\left(\frac{\ln a}{\omega}\right) - \Phi\left(-\frac{\ln a}{\omega}\right).$$

Como $\exists a$ tal que $F_T(a) = 0{,}5$, pode-se reescrever a expressão acima:

$$\Phi\left(\frac{\ln a}{\omega}\right) - \Phi\left(-\frac{\ln a}{\omega}\right) = 0{,}5 \qquad (3.2)$$

Sendo $v \sim N(0,1)$, portanto, simétrica, conclui-se que

$$\Phi\left(-\frac{\ln a}{\omega}\right) = P\left(v > \frac{\ln a}{\omega}\right), \text{ isto é}$$

$$\Phi\left(-\frac{\ln a}{\omega}\right) = 1 - P\left(v \leq \frac{\ln a}{\omega}\right)$$

$$\Phi\left(-\frac{\ln a}{\omega}\right) = 1 - \Phi\left(\frac{\ln a}{\omega}\right).$$

Substituindo esta última expressão na equação (3.2), obtém-se

$$\Phi\left(\frac{\ln a}{\omega}\right) - 1 + \Phi\left(\frac{\ln a}{\omega}\right) = 0{,}5.$$

Logo,

$$2 \cdot \Phi\left(\frac{\ln a}{\omega}\right) = \frac{3}{2} \quad \Longrightarrow \quad \Phi\left(\frac{\ln a}{\omega}\right) = \frac{3}{4}$$

$$\omega \cdot \Phi^{-1}\left(\frac{3}{4}\right) = \ln a \Longrightarrow a = \exp\left\{\omega \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\}.$$

Portanto,

$$MOR_{\text{aleat}} = \exp\left\{\omega \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\}. \text{ Como}$$

$$\omega^2 = \left(\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2}\right)\boldsymbol{\Sigma}\left(\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2}\right)^{\mathrm{T}} \quad \Longrightarrow \quad \omega^2 = 2\sigma^2.$$

Tem-se, então, que

$$MOR_{\text{aleat}} = \text{med}\{\exp(|\omega \cdot v|)\} = \exp\left\{\omega \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\}$$

$$\widehat{MOR}_{\text{aleat}} = \exp\left\{\sqrt{2\hat{\sigma}^2} \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\}$$

Resumindo, a razão de chances mediana relativa ao efeito aleatório representa a razão de chances entre os indivíduos de mais alto e mais baixo riscos, envolvidos com os mesmos valores de covariáveis para os efeitos fixos.

Utilizando o exemplo das pocilgas, tem-se que

$$\widehat{MOR}_{\text{aleat}} = \exp\left\{\sqrt{2 \cdot 1{,}38} \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\}$$

$$= \exp\left\{\sqrt{2 \cdot 1{,}38} \cdot 0{,}6745\right\}$$

$$\widehat{MOR}_{\text{aleat}} = 3{,}06$$

Interpretando este resultado, tem-se que entre pocilgas do mesmo tipo, o porco de mais alto risco tem três vezes mais chance de estar infectado que o porco de mais baixo

risco, o que, segundo Larsen et al. (2000) indica uma substancial heterogeneidade entre as pocilgas de mesmo tipo.

3.3.2 Razão de chances envolvendo o efeito fixo

Para quantificar o efeito fixo entre dois indivíduos escolhidos aleatoriamente, cada um dado por um padrão de variáveis explicativas, a razão de chances tem também a mesma forma apresentada na expressão (3.1):

$$OR = \exp(\kappa + \omega \cdot v)$$

Esta razão de chances é também uma variável aleatória e pelas mesmas razões apresentadas no item anterior, Larsen et al. (2000) propõem quantificar esta OR pela razão de chances mediana acrescida de um intervalo percentílico da variável aleatória V, $\text{perc}_a(V)$. Assim, esta razão de chances mediana representa a medida do efeito fixo entre duas unidades escolhidas aleatoriamente, cada uma com um determinado padrão de variáveis explicativas, e o intervalo percentílico reflete a variação da razão de chances devido ao efeito aleatório no preditor linear.

$$MOR_{\text{fixo}} = \text{med}\{\exp(\kappa + \omega \cdot v)\} = \exp(\kappa) \quad \text{e}$$

$$IOR = \left[\text{perc}_{(1-a)/2}\{\exp(\kappa + \omega \cdot v)\}; \text{perc}_{(1+a)/2}\{\exp(\kappa + \omega \cdot v)\}\right].$$

Esta razão de chances MOR_{fixo} refere-se ao objeto específico em estudo e o IOR caracteriza ambos o tamanho e a variação deste objeto. Note-se que o IOR não é um intervalo de confiança, é apenas uma ilustração das estimativas pontuais conjuntas de parâmetros fixos e aleatórios.

Os autores sugerem e utilizam em seus estudos um intervalo percentílico que abrange 80% da variação central da OR, isto é, $a = 0{,}8$. Logo,

$$\begin{aligned} IOR &= \left[\text{perc}_{(1-0{,}8)/2}\{\exp(\kappa + \omega \cdot v)\}; \text{perc}_{(1+0{,}8)/2}\{\exp(\kappa + \omega \cdot v)\}\right] \\ &= \left[\exp(\kappa + \omega \cdot \Phi^{-1}(0{,}10)) ; \exp(\kappa + \omega \cdot \Phi^{-1}(0{,}90))\right] \\ \widehat{\text{IOR}} &= \left[\exp\left(\kappa + \sqrt{2\hat{\sigma}^2} \cdot \Phi^{-1}(0{,}10)\right); \exp\left(\kappa + \sqrt{2\hat{\sigma}^2} \cdot \Phi^{-1}(0{,}90)\right)\right]. \end{aligned}$$

Voltando ao exemplo das pocilgas tem-se que

$$MOR_{\text{fixo}} = e^{\hat{\beta}_{\text{conv}} - \hat{\beta}_{\text{SPF}}} = 2{,}44$$

$$\widehat{\text{IOR}} = \left[\exp\left((-2{,}14 + 3{,}03) + \sqrt{2 \cdot 1{,}38} \cdot (-1{,}2816)\right);.\right.$$

$$\left.\exp\left((-2{,}14 + 3{,}03) + \sqrt{2 \cdot 1{,}38} \cdot (+1{,}2816)\right)\right]$$

$$\widehat{\mathrm{IOR}} = [0{,}291; 20{,}5]$$

A interpretação destes resultados é que, quando se comparam dois porcos tomados de duas pocilgas escolhidas aleatoriamente, a razão de chances irá, com uma probabilidade de 80%, encontrar-se dentro do intervalo do intervalo e, com probabilidade de 50%, será maior do que 2,44. Larsen et al. (2000) comentam que um intervalo desta amplitude sugere que seja importante melhorar a higiene nas pocilgas como um todo e a busca por outros fatores de risco poderia vir a ser mais frutífera.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A**. An introduction to categorical data analysis.** Hoboken: John Wiley & Sons, 2007.

AGRESTI, A. **Categorical data analysis**. Hoboken: John Wiley & Sons, 2002.

ASHFORD, J. R.; SOWDEN, R. R. Multvariate probit analysis. **Biometrics,** v. 26, p. 535-546, 1970.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares generalizados**. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA – SEAGRO, 12.; REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA – RBRAS, 52., 2007, Santa Maria. **Minicurso.** Santa Maria: UFSM, 2007.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica.** Piracicaba: ESALQ/USP, 2002.

FOX, J. **Applied regression analysis and generalized linear models.** 2.ed. Los Angeles: SAGE, 2008.

GARSON, D. **Logistic regression.** Syllabus for PA 766: Advanced Quantitative Research in Public Administration, 2010. Disponível em: <http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm>. Acesso em: 8 ago. 2010.

JACKMAN, S. **Generalized linear models**. Stanford University, 2011. Disponível em: < http://jackman.stanford.edu/papers/glm.pdf>. Acesso em: 3 mai. 2011.

KRAFT, I. G. G.; DE LEEUW, J. **Introducing multilevel modeling**. London: SAGE Puplications, 1998.

LARSEN, K. et al. Interpreting parameters in the logistic regression model with random effects. **Biometrics**, v. 56, n. 3, p. 909-914, 2000.

LARSEN, K.; MERLO, J. Appropriate assessment of neighborhood effects on individual health: interpreting random effects in multilevel logistic regression. **American Journal of Epidemiology**, Baltimore, v. 161, n. 1, p. 81-88, 2005.

LARSEN, K. New Measures for understanding the multilevel logistic regression model. In: WORKSHOP "STATISTICHE METHODEN FÜR KORRELIERTE DATEN", Bochum. **Seminário**. Bochum, 2006. Disponível em: <http://www.biometrie.uni-heidelberg.de/statmeth-ag/veranstaltungen/bochum06/Vortrag_Larsen.pdf>. Acesso em: 15 mai. 2012.

LINDSEY, J. K.; LAMBERT, P. On the appropriateness of marginal models for repeated measurements in clinical trials. **Statistics in Medicine**, v. 17, p. 447-469, 1998.

LITTELL, R. C.; FREUND, R. J.; SPECTOR, P. C. **SAS System for linear models.** Cary: SAS Institute Inc.,1991.

MERLO, J. et al. A brief conceptual tutorial of multilevel analysis in social epidemiologt: using measures of clustering in multilevel logistic regression do investigate contextual phenomena. **Journal of Epidemiology & Community Health**, v. 60, p. 290-297, 2006.

MERLO, J. Multilevel analytical approaches in social epidemiology: measures of health variation compared with traditional measures of association. **Journal of Epidemiology & Community Health**, v. 57, p. 550-552, 2003.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society,** v. 135, n. 3, p. 370-384, 1972.

OGLIARI. **Regressão logística.** Universidade Federal de Santa Catarina, [200-]. Disponível em: <www.inf.ufsc/~ogliari/arquivos/regressao_logistica.ppt>. Acesso em: 23 abr. 2010.

PAULA, G. **Modelos de regressão com apoio computacional**. IME/USP, São Paulo, 2004. Disponível em: <http://www.ime.usp.br/~giapaula/texto_2010.pdf>. Acesso em: 3 mai. 2011.

PINHEIRO, J. C.; BATES, D. M. **Mixed-effects models in S and S-Plus.** New York: Springer, 2000.

PIRES, F. O. et al. Característica sigmóide da FC durante teste progressivo e aplicação de diferentes métodos de identificação dos limiares de FC**. Revista Mackenzie de Educação Física e Esporte**, v. 7, n. 1, p. 45-58, 2008. Disponível em: <http:// www3.mackenzie.br/editora/index.php/remef/article/view/1213/904>. Acesso em: 8 dez. 2011.

REGAZZI, A. **EST 640 – Modelos lineares I.** [Apostila]. Departamento de Estatística, Universidade Federal de Viçosa. Viçosa: [s.n.], 2010.

RODRÍGUEZ, G. **Lecture notes on generalized linear model theory.** Princeton University, 2007. Disponível em: <http://data.princeton.edu/wws509/notes/a2.pdf>. Acesso em: 30 abr. 2011.

RUGGIERO, M. A. G.; LOPES, V. L. R. **Cálculo numérico: aspectos teóricos e computacionais.** São Paulo: Makron Books, 1996.

SANTOS, V. R. B. **Curso de cálculo numérico.** Rio de Janeiro: LTC, 1982.

SCIENTIFIC SOFTWARE INTERNATIONAL – SSI. **Generalized linear models**. [201-]. Disponível em: <http://www.ssi.com/lisrel/techdocs/sglim.pdf>. Acesso em: 5 mai. 2011.

SNIJDERS, T. A. B. Fixed and random effects. In: **Encyclopedia of Statistics in Behavioral Science.** v. 2, p. 664-665. Chicester: Wiley, 2005.

STIRATELLI, R.; LAIRD, N.; WARE, J. Random effects statistical model for serial observations with binary response. **Biometrics,** v. 40, p. 961-971, 1984.

VENEGAS, J. G.; HARRIS, R. S.; SIMON, B. A. A comprehensive equation for the pulmonary pressure-volume curve. **Journal of Applied Physiology**, Boston, v. 84, n. 1, p. 389-395, 1998.

WEBER. **Fixed and random effects.** University of California, San Diego [200-]. Disponível em: <http://weber.ucsd.edu/~tkousser/December%202%20Fixed%20and%20Random%20Effects.doc>. Acesso em: 3 abr. 2011.

WILLIAMS, D. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. **Biometrics**, v. 31, p. 949-952, 1975.

ZEGER, S. L.; LIANG, K.-Y.; ALBERT, P. S. Models for longitudinal data: a generalized estimating equation approach. **Biometrics**, v. 44, p. 1049-1060, 1988.

CAPÍTULO 4 - COMPARAÇÃO DE MÉTODOS DE ESTIMAÇÃO E PREDIÇÃO EM REGRESSÃO LOGÍSTICA COM EFEITO ALEATÓRIO

RESUMO: O objetivo deste trabalho foi comparar o desempenho de métodos de estimação por máxima verossimilhança para modelos de regressão logística com efeito aleatório introduzido no intercepto (máxima verossimilhança, máxima verossimilhança restrita e aproximação de Laplace), utilizando dados simulados e cenários definidos levando em consideração as características e a interpretação dos parâmetros. Pelo cálculo da porcentagem de simulações em que a variância do efeito aleatório se apresentou próxima de zero e da determinação do erro quadrático médio (EQM) foi verificada de forma indireta a acurácia de cada método para cada parâmetro envolvido no modelo. Concluiu-se que para a variância do efeito aleatório igual a 0,1, as estimativas obtidas usando aproximação de Laplace e ML consideraram que em torno de 35% os conjuntos de dados provinham de modelo com apenas efeito fixo, sendo REML mais robusto, considerando apenas 21,4%. As estimativas obtidas por Laplace apresentaram maior EQM para qualquer um dos parâmetros estimados, podendo ser considerado o pior procedimento. ML e REML se equivaleram em termos de EQM, porém, o último deve ser preferido por ser mais robusto em relação à variância zero. Por último, a variância do efeito aleatório influencia o EQM de um método, sendo diretamente proporcional a este.
PALAVRAS-CHAVE: aproximação de Laplace, erro quadrático médio, máxima verossimilhança, máxima verossimilhança restrita, modelos lineares generalizados mistos.

1 Introdução

Na análise de dados experimentais é comum a ocorrência de respostas binárias, cuja normalidade ou homocedasticidade dos erros não podem ser assumidas, sendo feita a modelagem da relação entre estas variáveis e variáveis explicativas por meio da regressão logística. Porém, sob este enfoque, se as observações forem correlacionadas, apresentando heterogeneidade dentro de níveis das variáveis explicativas, a incorporação de efeitos aleatórios ao modelo se faz necessária.

Os modelos de regressão logística com efeito aleatório pertencem à classe dos modelos lineares generalizados mistos (GLMM), cuja estimação de efeitos fixos e

componentes de variância e predição de efeitos aleatórios são realizadas simultaneamente, exigindo a utilização de métodos específicos para comportar tal complexidade.

Há métodos de estimação baseados nos quadrados mínimos (LS) como a ANOVA, na máxima verossimilhança (ML) e suas generalizações, como máxima quasiverosimilhança, REML, aproximação de Laplace dentre outros. De forma geral, a identificação daqueles mais eficientes é realizada via simulação de dados, considerando diversos cenários envolvendo diferentes graus de desbalanceamento e diferentes valores para os parâmetros e componentes de variância. Porém, na maioria dos trabalhos que envolvem simulação (PINHEIRO e BATES, 1995; RAUDENBUSH e YANG e YOSEF, 2000; CUSTÓDIO e BARBIN, 2005), os valores paramétricos são assumidos de forma empírica, não levando em consideração aspectos teóricos da regressão logística com efeito aleatório.

Tendo em vista a comparação de métodos de estimação para modelos não lineares mistos, como é o caso dos GLMM, Pinheiro e Bates (1995) via análise de dados reais e de simulação, relatam que a aproximação de Lapalace e quadratura Gaussiana apresentaram-se igualmente eficientes, enquanto o REML mostrou-se adequado apenas para parte dos modelos considerados. Com respeito ao modelo logístico com efeito aleatório, Raudenbush, Yang e Yosef (2000) compararam diferentes métodos baseados em ML e ML penalizada via simulação de dados sob diferentes cenários e concluíram que a aproximação de Laplace apresentou maior precisão e eficiência computacional. Utilizando este mesmo modelo, Custódio e Barbin (2005) também utilizaram técnicas de simulação e verificaram que os métodos de máxima quasiverossimilhança (QV) e máxima verossimilhança (LM) não apresentaram diferenças significativas entre si. Embora estes vários estudos tenham sido conduzidos tendo em vista a comparação de métodos de estimação, não há relatos de estudos que comparem especificamente os métodos de aproximação de Laplace, ML e REML e que apresentem aspectos computacionais detalhados em relação aos mesmos.

Diante do exposto, objetivou-se avaliar por meio de técnicas de simulação o desempenho de diferentes abordagens de estimação (aproximação de Laplace, ML e REML) considerando o modelo de regressão logística com efeito aleatório introduzido no intercepto. Além disso, objetivou-se também apresentar de forma detalhada a construção dos cenários de simulação levando em consideração a interpretação dos parâmetros e as características da regressão logística.

2 Material e métodos

2.1 Métodos de estimação para modelos lineares generalizados mistos

O modelo de regressão logística com efeito aleatório no intercepto (LARSEN, 2000) é dado por:

$$\pi_{ij} = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}, \quad \text{sendo } \eta_{ij} = (\beta_0 + u_j) + x\beta \text{ e } \pi_{ij} = P(y_{ij} = 1 | X = x) \quad (1)$$

cuja linearização leva a:

$$\text{logit}(\pi_{ij}) = ln\frac{\pi_{ij}}{1 - \pi_{ij}} = \eta_{ij}.$$

Nestas expressões tem-se que: η_{ij} é o preditor linear da *i*-ésima observação inserida na parcela *j*; x é o tratamento associado à observação: 0 para o tratamento referência e 1 para o tratamento com maior probabilidade de sucesso (tratamento 1); β_0 é o efeito associado ao tratamento referência; β é o efeito associado ao tratamento 1; u_j é o efeito aleatório associado à *j*-ésima parcela, com $u_j \sim N(0, \sigma_u^2)$. A equação do preditor linear pode ser reescrita em termos matriciais da seguinte forma:

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}, \quad (2)$$

em que: $\boldsymbol{\eta}_{n\times 1}$ é o vetor contendo os valores dos preditores lineares, $\boldsymbol{X}_{n\times p}$ é a matriz de incidência para os efeitos fixos, $\boldsymbol{\beta}_{p\times 1}$ é o vetor de de efeitos fixos (variáveis explicativas), $\mathbf{Z}_{n\times r}$ é a matriz de incidência para os efeitos aleatórios e $\mathbf{u}_{r\times 1}$ é o vetor de efeitos aleatórios, os quais seguem distribuição normal com média $\mathbf{0}$ e matriz de covariâncias $\boldsymbol{\Sigma} = \sigma_u^2\mathbf{I}$, sendo $\mathbf{I}$ uma matriz identidade.

As estimativas dos efeitos fixos ($\boldsymbol{\beta}$) e do componente de variância ($\boldsymbol{u}$), bem como as predições dos efeitos aleatórios (σ_u^2) do modelo apresentado em (2) podem ser obtidas considerando a formulação geral das equação dos modelos mistos dada por (RESENDE e BIELE, 2002):

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{X} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widetilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{y}^{*} \\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{y}^{*} \end{bmatrix}, \quad (3)$$

em que $\boldsymbol{S}^{-1}$ é matriz com termos diagonais dados por $\pi_{ij}(1 - \pi_{ij})\frac{1}{\sigma_{e_L}^2}$; $\sigma_{e_L}^2$ é variância residual na escala latente, igual a $\pi^2/3$ para a função de ligação logito (RESENDE, 2002); e $\boldsymbol{y}^{*} = g(\boldsymbol{y}) = g(\boldsymbol{\pi}) + (\boldsymbol{y} - \boldsymbol{\pi})g'(\boldsymbol{\pi})$.

Assim tem-se:

$$y_{ij}^{*} = \eta_{ij} + \frac{y_{ij} - \pi_{ij}}{\pi_{ij}(1 - \pi_{ij})}$$

$$= \log \frac{\pi_{ij}}{1-\pi_{ij}} + \frac{y_{ij}-\pi_{ij}}{\pi_{ij}(1-\pi_{ij})}.$$

Uma vez que o sistema apresentado em (3) é não linear quando se assume $\boldsymbol{\sigma}_u^2$ como desconhecida, métodos de estimação baseados em processos iterativos precisam ser utilizados.

O método da máxima verossimilhança (ML) consiste em maximizar a função densidade de probabilidade das observações em relação aos efeitos fixos e aos componentes de variância. O método é iterativo e fornece sempre estimativas não negativas das componentes de variância, desde que valores iniciais não negativos sejam dados. As estimativas são viesadas, pois o método não leva em conta a perda de graus de liberdade resultante da estimação dos efeitos fixos (PEREIRA e FERREIRA, 2008).

A aproximação de Laplace é usado como alternativa para determinar uma aproximação da função log-verossimilhança através de uma série de Taylor de segunda ordem, produzindo a log-verossimilhança perfilada aproximada (BATES, 2011). A avaliação de cada valor estimado dos parâmetros da log-verossimilhança perfilada requer a solução de um sistema não linear de quadrados mínimos penalizados, simultaneamente para ambos os parâmetros $\boldsymbol{u}$ e $\boldsymbol{\beta}$ do preditor linear.

O método da máxima verossimilhança restrita (REML) é uma modificação do método da ML proposta por Patterson e Thompson (1971), na qual a parte da função de verossimilhança que é invariante a $\boldsymbol{X\beta}$ é maximizada. Os estimadores REML levam em conta os graus de liberdade envolvidos nas estimativas dos efeitos fixos (CAMARINHA FILHO, 2003), eliminando o viés quando os dados são balanceados (SEARLE, 1987).

Para implementação computacional dos métodos citados, empregou-se o *software* R (R DEVELOPMENT CORE TEAM, 2011) por meio dos pacotes *lme4* (BATES e MAECHLER e BOLKER, 2011) e *nlme* (PINHEIRO et al., 2012). O primeiro foi utilizado para o método de aproximação de Laplace via função *lmer*; já o segundo foi utilizado para os métodos ML e REML.

A porcentagem de simulações por cenário nas quais $\sigma_u^2 \approx 0$ foi considerada um avaliador de qualidade dos métodos, sendo que quanto menor tal porcentagem, melhor o método, uma vez que simulou-se sob a hipótese de modelo aleatório ($\sigma_u^2 > 0$). Para avaliar o viés e a variância dos estimadores de forma conjunta, utilizou-se o Erro Quadrático Médio (EQM) considerando as estimativas dos parâmetros (β_0, β e σ_u^2) proporcionadas pelos diferentes métodos.

$$\mathrm{EQM}_{\theta} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta\right)^2}{n}.$$

2.2 Simulação

O estudo de simulação foi conduzido da seguinte forma: (i) atribuição de valores para β_0, β e σ_u^2, segundo definição detalhada dos cenários apresentada posteriormente; (ii) inclusão dos efeitos aleatórios (u_j) em cada parcela gerados de uma distribuição normal com média 0 e variância σ_u^2; (iii) determinação dos valores de η_{ij} e de π_{ij} conforme expressão (1); (iv) geração de uma variável aleatória auxiliar $k_{ij} \sim \text{Uniforme}[0,\ 1]$ para determinação da observação *y* a partir do seguinte critério: se $k_{ij} < \pi_{ij}$, então y = 1, caso contrário, y = 0; (v) análise dos conjuntos de dados gerados via funções do software R anteriormente especificadas para cada método utilizado, de tal forma que apenas foram consideradas análises válidas aquelas nas quais problemas de convergência (falsa convergência ou convergência singular) não foram detectados. Os códigos referentes aos itens (i) a (v) encontram-se disponibilizados no seguinte endereço eletrônico: http://www.det.ufv.br/~sebastiao/.

Foram realizadas 1.000 simulações para cada um dos 33 cenários considerados. Cada cenário simulava um experimento com delineamento inteiramente casualizado, composto de 2 tratamentos, 4 repetições com 50 observações cada uma, totalizando 8 parcelas (caracterizadas como efeitos aleatórios) (RESENDE e DUARTE, 2007) e 400 observações por cenário. O número de repetições e de observações por repetição foi baseado em um experimento (OLIVEIRA, 2009) para avaliar a germinação de sementes de pinhão-manso (*Jatropha curcas* L.).

2.3 Cenários

Conforme relatado anteriormente, os cenários foram definidos por meio de combinações de valores β_0, β e σ_u^2. Os valores para a variância do efeito aleatório, σ_u^2, foram definidos considerando graus diferentes de heterogeneidade entre as parcelas do experimento, a qual pode ser avaliada através da razão de chances mediana ($\mathrm{MOR}_{\mathrm{aleat}}$) (LARSEN et al., 2000). Esta representa a razão de chances entre os indivíduos de mais alto e mais baixo riscos assumindo os mesmos valores para os efeitos fixos, e é dada por:

$\widehat{MOR}_{\text{aleat}} = \exp\left\{\sqrt{2\hat{\sigma}_u^2} \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\} = \exp\left\{\sqrt{2\hat{\sigma}_u^2} \cdot 0{,}6745\right\}$, sendo $\Phi^{-1}\left(\frac{3}{4}\right)$ o valor resultante da inversa da função Normal acumulada avaliada em ¾. Assim, $\ln \widehat{MOR}_{\text{aleat}} = 0{,}6745\sqrt{2\hat{\sigma}_u^2}$, e $\hat{\sigma}_u^2 = \frac{(\ln \widehat{MOR}_{\text{aleat}})^2}{0{,}9099}$. Foram considerados os seguintes valores para σ_u^2 : 0,10, 1,00 e 2,00, correspondentes, respectivamente, aos valores 1,35, 2,60 e 3,85 de $MOR_{aleat.}$ Para maior facilidade de interpretação de tais valores, considerando um exemplo na área de germinação de sementes, a MOR_{aleat} = 1,35 representa que a melhor semente tem 1,35 vezes mais chance de germinação (aumento de 35% na chance de germinação) que a pior, dado que ambas compõem o mesmo tratamento, porém em parcelas diferentes. Assim, os valores escolhidos para a MOR_{aleat} representam diferentes graus de heterogeneidade entre as parcelas.

Os valores de β_0 foram especificados de acordo com a probabilidade de ocorrência de sucesso do tratamento referência (π_0), sendo $\beta_0 = \ln\frac{\pi_0}{1-\pi_0}$ e $\pi_0 = P(Y = 1|x = 0)$. Foram considerados os seguintes valores para β_0: -2,20, -0,85 e 0,41, correspondentes, respectivamente, aos valores 0,10, 0,30 e 0,60 de π_0.

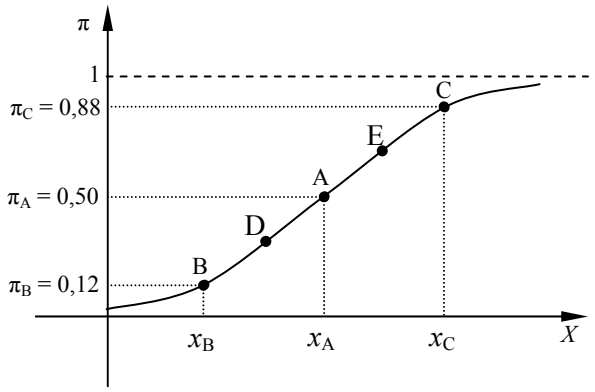


Figura 1 – Pontos relevantes da sigmoide: A é o nível mediano de efetividade, B e C são pontos de estabilização.

Para a definição de valores de β levou-se em consideração as características da curva de regressão logística, a sigmoide, mostradas na Figura 1. O ponto A representa o nível mediano de efetividade (AGRESTI, 2002), isto é, onde as probabilidades de sucesso e de fracasso são iguais ($\pi_A(x_A)$ = 0,50). Os pontos B e C são os pontos de estabilização ou deflexão (VENEGAS e HARRIS e SIMON, 1998), de forma que variações em x (variável explicativa) antes de x_B e depois de x_C acarretam variações ínfimas em relação a $\pi(x)$. Uma vez que os valores de π nos pontos A, B e C são, respectivamente, sempre 0,50, 0,12 e 0,88, assumiu-se também valores intermediários,

0,35 e 0,65, correspondentes aos pontos D e E. Tais valores de $\pi(x)$, juntamente com os valores já previamente estabelecidos de β_0, possibilitaram a obtenção dos valores de β (Tabela 1), por meio da expressão:

$\beta = \ln\frac{\pi_1}{1-\pi_1} - \beta_0$, onde $\pi_1 = P(Y = 1|x = 1)$. Estes valores de β refletem diferentes efeitos associados ao tratamento 1 (considerado como sendo aquele de maior probabilidade de sucesso).

Tabela 1 – Valores de β dados os valores de π_1 e de β_0

π_1	β_0		
	0,10	0,30	0,60
0,12	0,20	-	-
0,35	1,58	0,23	-
0,50	2,20	0,85	-
0,65	2,82	1,47	0,21
0,88	4,19	2,84	1,59

Associando-se aos valores de β e β_0 presentes na Tabela 1 aos valores de σ_u^2 iguais a 0,10, 1,00 e 2,00, foram obtidos os 33 cenários para simulação.

3 Resultados e discussão

Com o intuito de avaliar a capacidade dos métodos em produzir $\sigma_u^2 > 0$, foram calculadas as porcentagens de simulações por cenário nas quais $\sigma_u^2 \approx 0$ (Tabela 2). Os resultados mostraram que em média, ao simular assumindo $\sigma_u^2 = 0{,}1$, o método da aproximação de Laplace e o ML proporcionaram estimativas de variância $\sigma_u^2 \approx 0$ para 35,7 e 33,5% das simulações, respectivamente; enquanto que para o REML tal porcentagem foi significativamente menor, sendo esta 21,4%. Para os valores de σ_u^2 iguais a 1 e 2, em média todos os métodos apresentaram valores insignificantes para a referida porcentagem. Portanto, estes resultados indicam que o método REML é mais robusto em relação à estimação de componentes de variância iguais a zero ao se assumir modelos de regressão logística com efeito aleatório no intercepto.

Tendo em vista a grande quantidade de cenários contemplados, e a redundância dos resultados apresentados pelos mesmos, realizou-se uma filtragem a fim de se explorar apenas resultados provenientes de cenários contrastantes. Os cenários escolhidos são apresentados na Tabela 3.

Tabela 2 – Percentual de variâncias do efeito aleatório estimadas menores que 10^{-5}

$\sigma_u^2 = 0{,}10$				$\sigma_u^2 = 1{,}00$				$\sigma_u^2 = 2{,}00$			
Cenário	Laplace	ML	REML	Cenário	Laplace	ML	REML	Cenário	Laplace	ML	REML
C1	48,1	49,5	30,4	C2	3,1	3,7	1,9	C3	1,2	1,4	1,8
C4	38	30,4	19,5	C5	2	2,7	1,6	C6	0,8	1,2	0,8
C7	37,9	29,1	20,4	C8	2,2	2	0,9	C9	0,5	0,4	0,2
C10	37	29,2	20,1	C11	1,2	2,5	1,2	C12	0,4	0,8	0,3
C13	49,2	50,5	31,9	C14	3,5	3,6	2	C15	2,1	2,1	1,1
C16	35,7	32	18,7	C17	0,8	0,9	0,5	C18	0,2	0,2	0,1
C19	15,8	27,7	15,8	C20	0,6	0,8	0,3	C21	0	0,1	0
C22	30	30,3	18,4	C23	0,7	0,9	0,5	C24	0,1	0,1	0,1
C25	36,1	31,9	20,7	C26	1,6	1,8	0,9	C27	0,3	0,4	0,1
C28	28	28,3	17,7	C29	0,9	1	0,3	C30	0,1	0,1	0
C31	36,9	29,9	21,7	C32	1,5	1,8	1,1	C33	0,6	0,9	0,5
Média	35,7	33,5	21,4		1,6	2,0	1,0		0,6	0,7	0,5

Tabela 3 – Cenários selecionados

Cenários	β_0	β	σ_u^2	π_0	π_1
C4	-2,20	1,58	0,10		
C5	-2,20	1,58	1,00	0,10	0,35
C6	-2,20	1,58	2,00		
C19	-0,85	0,85	0,10		
C20	-0,85	0,85	1,00	0,30	0,50
C21	-0,85	0,85	2,00		
C31	0,41	1,59	0,10		
C32	0,41	1,59	1,00	0,60	0,88
C33	0,41	1,59	2,00		

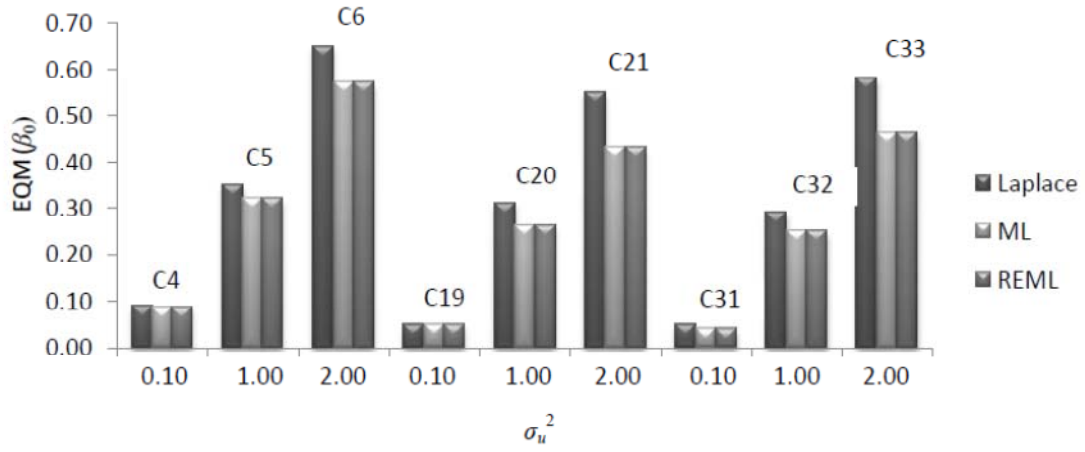
Na Figura 2 são apresentados os valores de EQM produzidos pelos métodos utilizados considerando cada um dos cenários da Tabela 3. Nota-se nesta Figura que a eficiência da estimação dos parâmetros β_0, β e σ_u^2 foi claramente influenciada pelos valores de σ_u^2, de forma que os valores de EQM aumentaram proporcionalmente ao aumento dos valores de σ_u^2. De forma geral, entende-se que quanto maior o valor de σ_u^2, maior a heterogeneidade dentro de níveis das variáveis explicativas, e realmente menos precisas serão as estimativas obtidas.

Nota-se ainda na referida Figura, que os métodos ML e REML superaram, em relação ao desempenho na estimação, o método de aproximação de Laplace. Esta

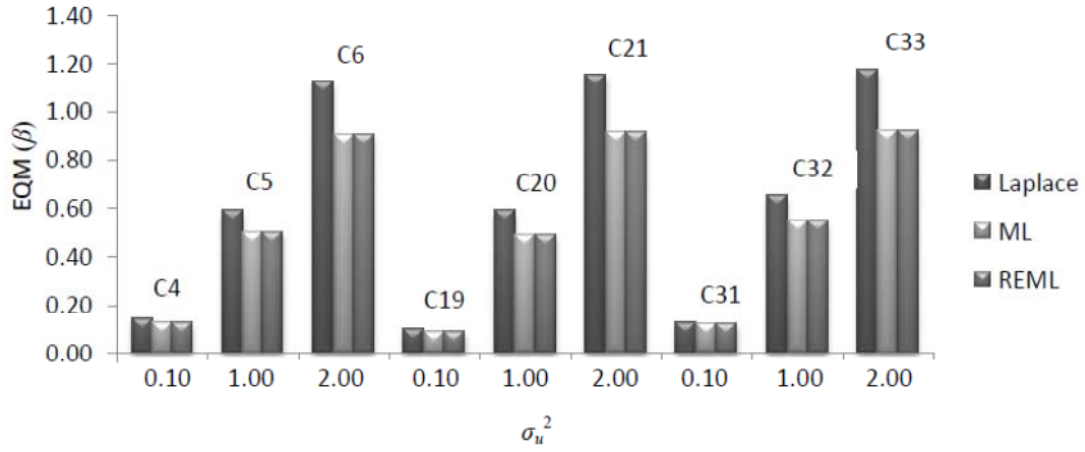
superioridade é diretamente proporcional ao aumento dos valores de σ_u^2, de forma que em situações experimentais com elevada heterogeneidade dentro de níveis das variáveis explicativas, o método de aproximação de Laplace deve ser evitado.

Tendo em vista que o método REML foi mais robusto em relação à estimação de componentes de variância iguais a zero (Tabela 1) e sua eficiência de estimação quanto aos valores de EQM na Figura 2, o mesmo foi considerado como sendo o método mais admissível nas condições do presente trabalho.

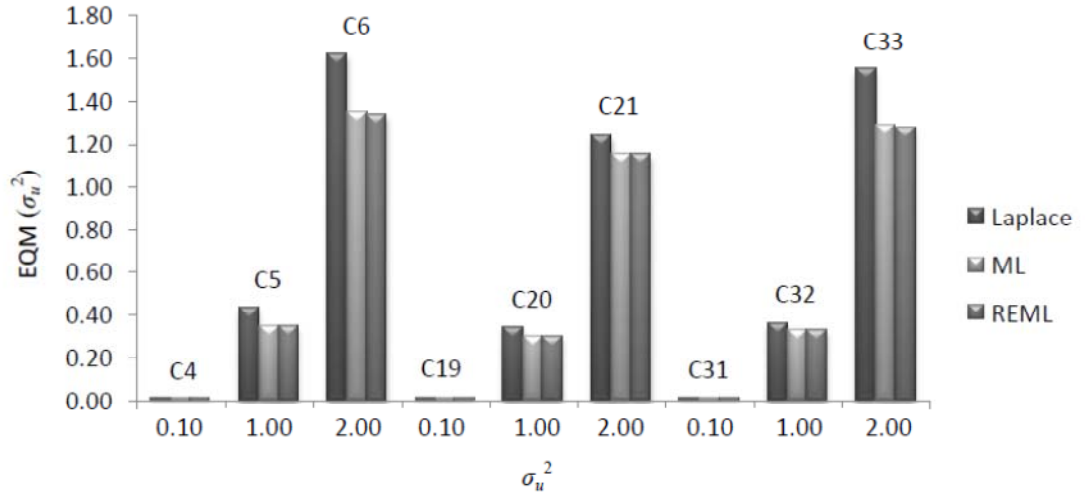
Até o presente momento, exaltou-se a relevância de σ_u^2 para se avaliar a eficiência dos métodos, porém, também é possível verificar resultados interessantes fixando valores de σ_u^2 e avaliando a relevância de combinações de β_0 e β (valores de π_0 e π_1 na Tabela 3) na avaliação dos métodos. Dessa forma, ao se considerar o maior valor de σ_u^2 (2,00) nota-se, especificamente nas Figuras 2(a) e 2(c), que os menores valores de EQM foram observados para o cenário C21 que contempla o valor $\pi_1 = 0{,}50$ (iguais probabilidades de sucesso e fracasso). Com relação a isto, Abdel-Azim e Berger (1999) e Garcia (2010) relataram que a qualidade de ajuste de GLMM para dados binários aumenta em função da diminuição das diferenças entre as probabilidades de sucesso e fracasso, e fazendo $\pi = 1 - \pi = 1/2$ obtém-se a melhor aproximação para uma distribuição normal.



(a)



(b)



(c)

Figura 2 – EQM para comparação entre os métodos da aproximação de Laplace, ML e REML: (a) para o intercepto; (b) para o parâmetro regressor; e (c) para a variância do efeito aleatório

Conclusões

Ao se assumir modelos de regressão logística com efeito aleatório no intercepto, nas simulações em que a variância do efeito aleatório foi igual a 0,1, o método da aproximação de Laplace e o ML tenderam a considerar que em torno de 35% dos conjuntos de dados possuíam apenas efeito fixo, sendo o método REML mais robusto em relação à estimação de componentes de variância diferentes de zero. Os métodos ML e REML apresentaram praticamente mesmo desempenho para todos os parâmetros estimados. Verificou-se que o desempenho na estimação do intercepto, do parâmetro regressor e da variância do efeito aleatório foram inversamente proporcionais à variância do efeito aleatório paramétrica. Ao se considerar a variância do efeito aleatório igual a 2,00, obteve-se melhor desempenho nas estimativas do intercepto e da variância do efeito aleatório quando as probabilidades de sucesso e fracasso do tratamento alternativo (tratamento 1) eram iguais.

Agradecimentos

Ao Programa de Pós-Graduação em Biometria e Estatística Aplicada do Departamento de Estatística da Universidade Federal de Viçosa e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

DUBOC, G.; MARTINS FILHO, S.; SILVA, F.F. Comparison of estimation and prediction methods in logistic mixed models.

ABSTRACT: The objective of this work was to compare the performance of estimation methods by maximum likelihood for logistic mixed models with random effect on intercept (maximum likelihood, restricted maximum likelihood and Laplace approximation), using simulated data and defined scenes considering characteristics and interpretation parameters. By calculating the percentage of simulations in which the variance of the random effects were close to zero and determining the mean square error (MSE) was measured indirectly the accuracy of each method for each parameter involved in the model. It was concluded that for the variance of the random effect equal to 0.1, the estimates obtained using the Laplace approximation and ML considered that about 35% the data sets came from model with only fixed effect, being REML more robust, considering just 21.4%. Estimates obtained by Laplace showed higher MSE for any parameters estimated, can be considered the worst procedure. ML and REML were

equivalent in terms of MSE, however, the latter should be preferred to be more robust with respect to zero variance. Finally, the variance of the random influences EQM of method, being directly proportional to this.

KEY WORDS: generalized linear mixed models, Laplace's method, maximum likelihood, mean squared error, restricted maximum likelihood.

Referências

ABDEL-AZIM, G. A.; BERGER, P. J. Properties of threshold model predictions. *Journal of Animal Science,* v.77, p.582-590, 1999.

AGRESTI, A. *Categorical data analysis.* Hoboken: John Wiley & Sons, 2002.

BATES, D. *Computational methods of mixed models*. University of Wisconsin, Department of Statistics. Madison, 2011. Disponível em: < http://cran.r-project.org/ web/packages/lme4/vignettes/Theory.pdf>. Acesso em: 8 out. 2011.

BATES, D.; MAECHLER, M.; BOLKER, B. *lme4: linear mixed-effects models using S4 classes*. version 0.999375-42, 2011. Disponível em: <http://lme4.r-forge.r-project.org/>. Acesso em: 12 out. 2011.

CAMARINHA FILHO, J.A. *Notas metodológicas sobre modelos lineares mistos***.** Departamento de Estatística, Universidade Federal do Paraná. Curitiba: 2003. Disponível em: < http://www.est.ufpr/rt/jom03a.pdf>. Acesso em: 8 jun. 2011.

CUSTÓDIO, T. N.; BARBIN, D. Comparação de modelos mistos visando à estimação do coeficiente de herdabilidade para dados de proporções. *Revista de Matemática e Estatística,* São Paulo, v.23, n.2, p.23-31, 2005.

GARCIA, D. A. *Avaliação genética da prenhez precoce em animais da raça Nelore utilizando modelos lineares generalizados mistos*. 2010. Dissertação (Mestrado em Zootecnia) – Universidade Federal dos Vales do Jequitinhonha e Mucuri, 2010.

LARSEN, K. et al. Interpreting parameters in the logistic regression model with random effects. *Biometrics,* v.56, n.3, p.909-914, 2000.

OLIVEIRA, G. L. *Testes para avaliação da qualidade fisiológica de sementes de pinhão-manso (Jatropha curcas L.)*. 2009. Dissertação (Mestrado em Fitotecnia) – Universidade Federal de Viçosa, Viçosa, MG, 2009.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block when blocks sizes are unequal. *Biometrika***,** v.58, n.3, p.545-554, 1971.

PEREIRA, L.; FERREIRA, L. Estimação de modelos lineares gerais mistos utilizando o SAS®. *Revista da Escola Superior de Gestão, Hotelaria e Turismo da Universidade do Algarve,* Faro, v.7, n.17, p.44-51, 2008. Disponível em: <http:// www. dosalgarves. com/revista/N17/7rev17.pdf>. Acesso em: 23 out. 2011.

PINHEIRO, J. C.; BATES, D. M. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics,* v.4, n.1, p. 12-35, 1995.

PINHEIRO, J. C.et al. *nlme: linear and nonlinear mixed effects models*. R package version 3.1-103, 2012.

R DEVELOPMENT CORE TEAM, *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0, URL: http://www.R-project.org/.

RAUDENBUSH, S. W.; YANG, M.-L.; YOSEF, M. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation: *Journal of Computational and Graphical Statistics*, v.9, n. 1, p.141-157, 2000.

RESENDE, M. D. V.; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. *Revista de Matemática e Estatística*, v.20, p.30-65, 2002.

RESENDE, M. D. V.; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesquisa Agropecuária Tropical, v.37, n.3,p.182-194, 2007.

RESENDE, M. D. V. *Genética biométrica e estatística no melhoramento de plantas perenes*. Brasília: Embrapa Informação Tecnológica, 2002.

SEARLE, S. R. *Linear models for unbalanced data*. New York: John Wiley, 1987.

VENEGAS, J, G.; HARRIS, R. S.; SIMON, B. A. A comprehensive equation for the pulmonary pressure-volume curve. *Journal of Applied Physiology*, Boston, v.84, n.1, p.389-395, 1998.

CAPÍTULO 5 – REGRESSÃO LOGÍSTICA COM EFEITO ALEATÓRIO APLICADA NA GERMINAÇÃO DE SEMENTES DE PINHÃO-MANSO

MIXED LOGISTIC REGRESSION APPLIED ON SEED GERMINATION OF PHYSIC NUT

RESUMO

O objetivo deste trabalho foi ajustar um modelo de regressão logística com efeito aleatório no intercepto aos dados provenientes de um experimento de germinação de pinhão-manso (*Jatropha curcas* L.), envolvendo os substratos rolo de papel, sobre papel, entre areia e sobre areia, e classificar os substratos pela a razão de chances mediana, fazendo um paralelo com o resultado obtido pelo teste de Tukey para comparação de médias. Concluiu-se que a chance de germinação do substrato entre areia (EA), utilizando a razão de chances mediana, se mostrou superior ao sobre papel (SP), porém para outras combinações de substratos, esta se apresentou inconclusiva, resultado este também obtido a partir do teste de Tukey. A razão de chances para o efeito aleatório apontou considerável heterogeneidade de germinação de sementes em unidades diferentes de um mesmo substrato. Tal heterogeneidade comprometeu as análises inferenciais para o experimento.

Palavras-chave: intervalo percentílico, *Jatropha curcas* L., máxima verossimilhança restrita, pinhão-manso, razão de chances mediana, teste de Tukey.

ABSTRACT

The objective of this work was to adjust a logistic mixed model with random effect on intercept for data from an experiment of germination of physic nut (*Jatropha curcas* L.), involving the substrate paper roll, on paper, between sand and on sand, and

classify the substrates by the median odds ratio, making a parallel with the result obtained by the Tukey's test for comparison means. It was concluded that the chance of germination of between sand substrate (BS), using the odds ratio median, was statistically better than on paper (OP), but for other combinations of substrates, it appeared inconclusive, similarly result obtained from the Tukey's test. The odds ratio for the random effect showed considerable heterogeneity of seed germination in different units of the same substrate. Such heterogeneity compromised the inferential analyzes for the experiment.

Key words: *Jatropha curcas* L., median odds ratio, percentile interval, physic nut, restricted maximum likelihood, Tukey's test.

INTRODUÇÃO

O pinhão-manso (*Jatropha curcas* L.) é uma planta da família das *Euphorbiaceae*, que se destacou por ser uma oleaginosa com características ideais para a obtenção do biodiesel (CARNIELLI, 2003) e na reestruturação de área degradadas. O cultivo em larga escala do pinhão-manso pode ser limitado pela baixa germinação das sementes e heterogeneidade dos genótipos (CARVALHO, 2010). Experimentos em laboratório envolvendo diversos substratos de germinação procuram estudar a influência tanto na porcentagem de germinação quanto na determinação do vigor das plântulas (BEWLEY & BLACK, 1994). Sob este aspecto, a análise estatística desses dados constitui importante ferramenta para o desenvolvimento do estudo, sendo a regressão logística um dos principais modelos utilizados. Considerando a heterogeneidade germinativa de suas sementes e de outras condições experimentais, é plausível pensar numa modelagem dos dados com inclusão de efeito aleatório.

Muitos estudos lançam mão do teste de Tukey para comparação de médias, como no trabalho de SILVA et al. (2009) que visa conhecer o grau de sensibilidade do pinhão-manso em relação a diversos tipos de salinidade; no trabalho de NUNES et al (2008) que avalia a influência do estado de maturação dos frutos no desenvolvimento de embriões de pinhão-manso; e no de VANZOLINI et al. (2010) que estudou o efeito da temperatura e do tempo de contagem na germinação de sementes de pinhão-manso. Porém, como alertam CALDAS & FERNANDES (2011), procedimentos para comparações múltiplas de média só são válidos sob fortes pressuposições, como o caso do conhecido teste de Tukey que exige normalidade dos dados.

Dado o exposto, o objetivo deste trabalho foi ajustar um modelo de regressão logística com efeito aleatório no intercepto aos dados provenientes de um experimento de germinação de pinhão-manso (*Jatropha curcas* L.), para verificar o desempenho da razão de chances mediana na classificação dos substratos, através da comparação com o resultado obtido pelo teste de Tukey para comparação de médias.

Uma característica importante da regressão logística é a possibilidade de interpretação de seus parâmetros por meio da razão de chances (OR), a qual permite estabelecer relações e fazer predições a respeito dos tratamentos. No entanto, na presença de efeito aleatório, os parâmetros de efeito fixo não mantêm suas interpretações. Partindo do princípio que a OR, em presença de efeito aleatório, é uma variável aleatória, LARSEN et al. (2000) propõem a razão de chances mediana (MOR). Considerando, então, o modelo logístico misto

$$\pi_i = P(Y_i = 1|\eta_i) = \frac{\exp\{\eta_i\}}{1+\exp\{\eta_i\}} \text{ e } \text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i} = \eta_i \qquad (1)$$

tem-se $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u}$, no qual $\boldsymbol{u} \sim \text{N}(0,\ \sigma_u^2\mathbf{I})$ é o vetor de efeitos aleatórios e **I** uma matriz identidade. Sendo $\text{chance}(Y_i = 1|\boldsymbol{X} = \boldsymbol{x}_i) = \frac{\pi_i}{1-\pi_i} = e^{\eta_i}$, então a razão de chances (OR) entre duas medidas i_1 e i_2 é dada por $\text{OR} = \exp\left(\eta_{i_1} - \eta_{i_2}\right)$, que para o

modelo com efeito aleatório assume a forma

$$\text{OR} = \exp\{(\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2})\boldsymbol{\beta} + (\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2})\mathbf{u}\} = \exp(\kappa + \omega \cdot v), \qquad (2)$$

sendo $\kappa = (\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2})\boldsymbol{\beta}$, $\omega^2 = (\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2})(\sigma_u^2\mathbf{I})(\boldsymbol{z}_{i_1} - \boldsymbol{z}_{i_2})^{\mathrm{T}}$ e $v \sim N(0,1)$, cuja estimação e interpretação depende do efeito considerado, fixo ou aleatório.

Para a razão de chances envolvendo o efeito aleatório, LARSEN et al., (2000) propõem o uso da razão de chances mediana. Considerando dois indivíduos com os mesmos valores de explicativas para os efeitos fixos, a razão de chances mediana entre a unidade de mais alto risco (ou chance) e a de mais baixo risco, utilizando a expressão em (2), é dada por:

$$\text{OR} = \exp(\omega \cdot v)$$

$$MOR_{\text{aleat}} = \text{med}\{\exp(|\omega \cdot v|)\} = \exp\left\{\omega \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\},$$

cujo estimador é:

$$\widehat{\text{MOR}}_{\text{aleat}} = \exp\left\{\sqrt{2\hat{\sigma}_u^2} \cdot \Phi^{-1}\left(\frac{3}{4}\right)\right\}, \qquad (3)$$

em que $\hat{\sigma}_u^2$ é a variância estimada para o efeito aleatório.

Para quantificar os efeitos fixos entre dois indivíduos escolhidos aleatoriamente, cada um dado por um padrão de variáveis explicativas, os mesmos autores propõem o uso razão de chances mediana para este efeito acrescida de um intervalo percentílico (IOR) da variável aleatória V, $\text{perc}_a(V)$, o qual reflete a variação da razão de chances devido ao efeito aleatório no preditor linear. Desta forma,

$$\text{MOR}_{\text{fixo}} = \text{med}\{\exp(\kappa + \omega \cdot v)\} = \exp(\kappa) \quad \text{e}$$

$$\text{IOR} = \left[\text{perc}_{(1-a)/2}\{\exp(\kappa + \omega \cdot v)\}; \text{perc}_{(1+a)/2}\{\exp(\kappa + \omega \cdot v)\}\right]. \qquad (4)$$

O IOR não é um intervalo de confiança, é apenas uma ilustração das estimativas pontuais conjuntas de parâmetros fixos e aleatórios. LARSEN et al. (2000) e LARSEN & MERLO (2005) sugerem e utilizam em seus estudos um intervalo percentílico que

abrange 80% da variação central da OR, isto é, a = 0,8. Neste caso o IOR é estimado por

$$\widehat{\text{IOR}} = \left[\exp\left(\kappa + \sqrt{2\hat{\sigma}^2} \cdot \Phi^{-1}(0{,}10)\right); \exp\left(\kappa + \sqrt{2\hat{\sigma}^2} \cdot \Phi^{-1}(0{,}90)\right)\right]. \quad (5)$$

MATERIAL E MÉTODOS

Para avaliar a germinação de sementes de pinhão-manso (*Jatropha curcas* L.), foram utilizadas oito amostras de 50 sementes para cada um dos substratos: 1) rolo de papel (RP), no qual as sementes foram distribuídas sobre duas folhas e cobertas por uma folha de papel toalha umedecidas com três vezes o peso do substrato; 2) sobre papel (SP), onde as sementes foram colocadas sobre duas folhas de papel toalha em caixas do tipo gerbox e umedecidas com um volume de água referente a três vezes o peso do substrato seco; 3) sobre areia (SA) onde as sementes foram semeadas em caixas plásticas tipo gerbox, preenchidas com 118 ml de areia e sobre a qual foram depositadas as sementes e o substrato foi umedecido até atingir a capacidade de campo; 4) entre areia (EA) onde se seguiu a mesma metodologia descrita para o substrato SA, sendo que as sementes foram cobertas por uma camada de 1,5 cm de areia. A contagem das plântulas normais foi feita aos 10 dias após a instalação dos testes.

O experimento foi conduzido em condições de laboratório na Universidade Federal de Viçosa, utilizando o delineamento inteiramente casualizado, com 8 repetições de 50 sementes, procedendo-se desta forma todas as análises necessárias dos dados.

Para descrever a probabilidade de germinação das sementes foi usado um modelo logístico na presença de efeito aleatório. As sementes de uma mesma unidade de substrato foram consideradas correlacionadas entre si, levando à modelagem de efeito aleatório para o substrato. No experimento em questão tem-se que: n = 1.600 (número total de observações), p = 4 (número de parâmetros para os efeitos fixos), r = 32

(número de parâmetros para os efitos aleatórios = número total de unidades de substratos (parcelas) utilizados no experimento), e $\boldsymbol{u} \sim N(0,\ \sigma_u^2)$ é vetor de efeitos aleatórios não mensuráveis usado para modelar a correlação entre as unidades de uma mesma parcela. A variável explicativa tipo de substrato, por ser qualitativa com mais de duas categorias, foi reparametrizada criando-se três variáveis indicadoras (variáveis *dummy*), o SP foi recodificado para *Sd*1, o SA para *Sd*2 e o EA para *Sd*3 e o RP como variável de referência. O trio ordenado (*Sd*1, *Sd*2, *Sd*3) indica o substrato usado: (0, 0, 0) refere-se ao substrato referência RP, (1, 0, 0) ao substrato SP, (0, 1, 0) ao substrato SA e (0, 0, 1) ao EA. Nestas condições, o modelo logístico dado pela expressão (1) ficou definido como

$$\text{logit}(\pi_i) = \eta_i = \beta_0 + \beta_1(Sd_1) + \beta_2(Sd_2) + \beta_3(Sd_3) + u_j, \qquad (6)$$

sendo a i-ésima observação sujeita ao efeito aleatório da j-ésima parcela.

Procedeu-se à análise do experimento fazendo uso do método da máxima verossimilhança restrita (REML), que produz estimativas não viesadas para dados balanceados (SEARLE, 1987). Para implementação computacional do método citado, empregou-se o *software* R (R DEVELOPMENT CORE TEAM, 2011) por meio pacote *nlme* (PINHEIRO et al., 2012). Depois foram calculadas e interpretadas as razões de chances medianas e o intervalo percentílico. Posteriormente verificou-se a normalidade dos dados por meio do teste de normalidade de Shapiro-Wilk a 5% de significância e aplicou-se o teste de Tukey aos dados.

RESULTADOS E DISCUSSÃO

Para o modelo descrito no item anterior (expressão (6)), fazendo uso do método REML, foram obtidas as estimativas apresentadas na Tabela 1.

Na comparação entre os tratamentos pela razão de chances mediana para os efeitos aleatório e fixos, foi utilizado o intervalo percentílico de 80%, conforme

sugestão de LARSEN et al. (2000) e LARSEN & MERLO (2005). Para o cálculo das estimativas de $\mathrm{MOR}_{\mathrm{aleat}}$, $\mathrm{MOR}_{\mathrm{fixo}}$ e IOR foram usadas, respectivamente, as expressões (3), (4) e (5), considerando: $\hat{\sigma}_u^2 = 0{,}246315$; $\Phi^{-1}(3/4) = 0{,}6745$; $\Phi^{-1}(0{,}10) = -1{,}2816$ e $\Phi^{-1}(0{,}90) = 1{,}2816$.

Os resultados estão apresentados na Tabela 2.

A razão de chances mediana do efeito aleatório mostrou que a melhor semente tem em torno de 61% mais chance de germinar que a pior, considerando unidades diferentes do mesmo substrato. Parte desta heterogeneidade pode ser atribuída à heterogeneidade inerente às sementes (CARVALHO, 2010) e parte aos procedimentos experimentais. Para determinar essas componentes seria preciso testar um novo modelo de efeitos mistos.

Quando comparados os substratos sobre papel e rolo de papel, verificou-se que medianamente sobre papel possui chance de germinação 55% menor que rolo de papel. Porém, ao ser observado o IOR, há uma probabilidade de 80% (dada pelo percentil do IOR) de que a chance de germinação de uma semente no substrato sobre papel possa ocorrer de 88% menos a 35% mais que uma outra no substrato rolo de papel. Este resultado mostrou que a comparação entre tais substratos é inconclusiva. Quando dois substratos de tipos diferentes apresentam iguais chances de germinação para suas sementes, a razão de chances é igual a 1. Assim, se um IOR contém o 1, este torna a razão de chances inconclusiva, pois a chance de germinação tanto pode diminuir como aumentar. Assim, pela Tabela 2, a única afirmação garantida é que a chance de germinação no substrato entre areia se mostrou superior ao sobre papel, podendo ser de 1,20 a 7,26 vezes melhor, sendo, medianamente igual a 2,95 vezes.

Percebeu-se, ainda que sendo $\hat{\sigma}_u^2 \approx 0{,}25$, a amplitude dos IOR's nas diversas comparações variou de 1,13 a 8,46 unidades como consequência da inclusão do efeito

aleatório no modelo, gerando repercussões na interpretação da razão de chances e provocando dúvidas sobre qualidade do experimento.

A normalidade dos dados foi confirmada pelo teste de Shapiro-Wilk a 5% de significância, sendo, então, realizado o teste de Tukey para os dados, considerando a mesma significância, cujos resultados estão apresentados na Tabela 3. Os valores apresentados pelo teste de Tukey mostram que a 5% de significância só há diferença significativa entre as médias dos tratamentos entre areia e sobre papel, sendo o primeiro superior ao segundo. Este resultado confirma aqueles obtidos por razão de chances mediana.

CONCLUSÃO

A razão de chances mediana associada ao intervalo percentílico mostrou ter bom desempenho na classificação dos substratos (tratamentos) em relação à germinação de sementes de pinhão-manso, sendo semelhante à comparação feita pelo teste de Tukey. Por não necessitar do pressuposto de normalidade dos dados e por ser possível a interpretação dos parâmetros, a razão de chances mediana para modelos de regressão logística com efeito aleatório incidindo no intercepto se torna uma ferramenta poderosa de análise.

AGRADECIMENTOS

Ao Programa de Pós-Graduação em Biometria e Estatística Aplicada do Departamento de Estatística da Universidade Federal de Viçosa e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERÊNCIAS

BEWLEY, J.D.; BLACK, M. **Seeds: physiology of development and germination**. New York: Plenum Press, 1994. 445p.

CALDAS, G. B.; FERNANDES, G. B. Comparações múltiplas de médias em modelos paramétricos através de contrastes gerais, restrições lógicas e correlações. In: ESCOLA DE MODELOS DE REGRESSÃO, 12, 2011, Fortaleza. **Anais**... Fortaleza: Universidade Federal do Ceará, 2011.

CARNIELLI, F.O combustível do futuro. **Boletim Informativo,** n.1413, 2003. Disponível em: <http://www.ufmg.br/boletim/bol1413/quarta.shtml>. Acesso em: 08 ago. 2010.

CARVALHO, J. M. F. C. et al. Avaliação de meios de cultivo para pinhão manso (*Jatropha curcas* L.). In: CONGRESSO BRASILEIRO DE MAMONA, 4; SIMPÓSIO INTERNACIONAL DE OLEAGINOSAS ENERGÉTICAS, 1, 2010, João Pessoa. **Anais**... Campina grande: Embrapa Algodão, 2010. p. 211-216.

LARSEN, K.; MERLO, J. Appropriate assessment of neighborhood effects on individual health: interpreting random effects in multilevel logistic regression. **American Journal of Epidemiology**, Baltimore, v.161, n.1, p.81-88, 2005.

LARSEN, K. et al. Interpreting parameters in the logistic regression model with random effects. **Biometrics**, Arlington, v.56, n.3, p.909-914, 2000.

NUNES, C. F. et al. Diferentes suplementos no cultivo in vitro de embriões de pinhão-manso. Pesquisa Agropecuária Brasileira, Brasília, v.43, n.1, p.9-14, 2008.

PINHEIRO, J.C.et al. **nlme: linear and nonlinear mixed effects models**. R package version 3.1-103, 2012.

R DEVELOPMENT CORE TEAM, *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0, URL: http://www.R-project.org/.

SEARLE, S. R. *Linear models for unbalanced data*. New York: John Wiley, 1987.

SILVA, E. N. et al. Acúmulo de íons e crescimento de pinhão-manso sob diferentes níveis de salinidade. **Revista Ciência Agronômica**, Fortaleza, v.40, n.2, p.240-246, 2009.

VANZOLINI, S. Qualidade sanitária e germinação de sementes de pinhão-manso. **Revista Brasileira de Sementes**, v.32, n.4, p.9-14, 2010.

Tabela 1 – Estimativas obtidas pelo método da máxima verossimilhança restrita (REML)

Parâmetro	Estimativa	Desvio padrão	G.L.	t-value	p-value
Intercepto: β_0	1,466222	0,223013	28	6,574596	0
β_1	-0,60132	0,299596	28	-2,00712	0,0545
β_2	-0,15788	0,30831	28	-0,51208	0,6126
β_3	0,481481	0,336816	28	1,429506	0,1639
Efeito Aleatório: σ_u^2	0,246315				

Tabela 2 – Razão de chances mediana para o efeito fixo e para o efeito aleatório

MÔR aleat	1,61	
Comparação	MOR fixo	IÔR
SP vs RP	0,55	[0,22; 1,35]
SA vs RP	0,85	[0,35; 2,10]
EA vs RP	1,62	[0,66; 3,98]
SA vs SP	1,56	[0,63; 3,83]
EA vs SP	2,95	[1,20; 7,26]
EA vs SA	1,90	[0,77; 4,66]

Tabela 3 – Comparação das médias dos tratamentos pelo teste de Tukey

Tratamento	Classificação	Média
EA	a	88
RP	ab	81,5
SA	ab	78,75
SP	b	70,25

CONCLUSÃO GERAL

Considerando o modelo de regressão logística com efeito aleatório no intercepto, os métodos da aproximação de Laplace, máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) apresentaram desempenho diferentes em relação aos dados simulados, tendo o primeiro melhor desempenho na estimativa dos parâmetros. A variância do efeito aleatório interfere no desempenho de qualquer um dos métodos avaliados, sendo variância e desempenho inversamente proporcionais. No experimento de avaliação de germinação de sementes de pinhão-manso (*Jatropha curcas* L.) envolvendo os substratos rolo de papel, sobre papel, sobre areia e entre areia, a razão de chances mediana para o efeito aleatório apontou considerável heterogeneidade na germinação de sementes em unidades diferentes de um mesmo substrato. A razão de chances mediana para o efeito fixo associada ao intervalo percentílico apresentou bom desempenho na classificação dos substratos, levando a resultados semelhantes aos do teste de Tukey. O substrato entre areia foi considerado melhor que sobre papel na germinação de sementes por ambas as técnicas.

APÊNDICE A – Códigos de programação no *software* R, versão 2.13.2 , usados no capítulo 4

```
# Programa para simular e ajustar modelos para dados binários com efeito aleatório


# Experimento com efeito aleatório nas parcelas e incidindo no intercepto
#   ntrat:      número de tratamentos
#   nrep:       número de repetições
#   nobs:       número de observações por repetição
#   tobs:       total de observações por experimento
#   nparcela:   total de parcelas por experimento (= número de efeitos aleatórios)
#   nsimul:     número de simulações

# Modelo: pi= exp{eta}/(1 + exp{eta}), onde eta = (beta0 + efeAle) + beta*x
#   pi:         matriz de probabilidade
#   eta:        matriz de efeitos fixos + efeitos aleatórios
#   beta0:      efeito fixo do tratamento referência (tratamento 0)
#   beta:       efeito fixo do tratamento 1
#   efeAle:     matriz de efeitos aleatórios
#   varEfAle:   variância do efeito aleatório
#   u:          vetor de efeitos aleatórios nas parcelas
#   intercp:    matriz de intercepto com efeito aleatório
#   k:          matriz de valores gerados de uma distribuição uniforme [0, 1]

# Conjunto de dados
#   dados:      banco de dados
#   simulação: vetor com o número da simulação
#   parcela:    vetor com o número da parcela
#   rep:        vetor com o número da repetição
#   x:          vetor com o número do tratamento
#   y:          vetor de observações geradas por simulação

# Ajuste de modelos

# A) Modelo de efeito fixo
#    fit:           ajuste do modelo
#    result_fixo: matriz de parâmetros estimados para cada simulação: intercepto (beta0)
                    e coeficiente (beta)

# B) Modelo de efeito aleatório: método de Laplace
#    fit1:          ajuste do modelo
#    int_ale1:      vetor com o intercepto estimado (beta0) para cada simulação
#    cor_ale1:      vetor de correlação entre os efeitos aleatórios simulados e estimados
                    para cada simulação
#    x_ale1:        vetor de coeficiente (beta) estimado para cada simulação
#    var_ale1:      variância do efeito aleatório estimada para cada simulação
#    efeAle_est: matriz contendo os efeitos aleatórios estimados de cada parcela, para
                    cada simulação
```

```
#   result_ale1: matriz contendo o resumo das estimativas de cada simulação

# C) Modelo de efeito aleatório: método da máxima verossimilhança (ML)
#   fit2:          ajuste do modelo
#   int_ale2:      vetor com o intercepto estimado (beta0) para cada simulação
#   cor_ale2:      vetor de correlação entre os efeitos aleatórios simulados e estimados
                   para cada simulação
#   x_ale2:        vetor de coeficiente (beta) estimado para cada simulação
#   var_ale2:      variância do efeito aleatório estimada para cada simulação
#   efeAle_est:    matriz contendo os efeitos aleatórios estimados de cada parcela, para
                   cada simulação
#   result_ale2: matriz contendo o resumo das estimativas de cada simulação

# D) Modelo de efeito aleatório: método da máxima verossimilhança restrita (REML)
#   fit3:          ajuste do modelo
#   int_ale3:      vetor com o intercepto estimado (beta0) para cada simulação
#   cor_ale3:      vetor de correlação entre os efeitos aleatórios simulados e estimados
                   para cada simulação
#   x_ale3:        vetor de coeficiente (beta) estimado para cada simulação
#   var_ale3:      variância do efeito aleatório estimada para cada simulação
#   efeAle_est:    matriz contendo os efeitos aleatórios estimados de cada parcela, para
                   cada simulação
#   result_ale3: matriz contendo o resumo das estimativas de cada simulação

# Concatenação dos dados de 100 simulações do laço
#   rfixo: resumo das estimativas do modelo fixo
#   rale1: resumo das estimativas do modelo misto por Laplace
#   rale2: resumo das estimativas do modelo misto por ML
#   rale3: resumo das estimativas do modelo misto por REML

# Concatenação dos dados das 1000 simulações
#   rfixomil: resumo das estimativas do modelo fixo
#   rale1mil: resumo das estimativas do modelo misto por Laplace
#   rale2mil: resumo das estimativas do modelo misto por ML
#   rale3mil: resumo das estimativas do modelo misto por REML


# Médias das estimativas obtidas nas simulações
#   médias_fixo: vetor das médias obtidas no modelo de efeito fixo
#   médias_ale1: vetor das médias obtidas no modelo misto por Laplace
#   médias_ale2: vetor das médias obtidas no modelo misto por ML
#   médias_ale3: vetor das médias obtidas no modelo misto por REML

# Erro quadrático médio

#   eqm_fixo_b0:  EQM do método de efeito fixo para o intercepto
#   eqm_fixo_b:    EQM do método de efeito fixo para o efeito do tratamento 1
#   eqm_ale1_b0:  EQM do método de Laplace para o intercepto
#   eqm_ale1_b:    EQM do método de Laplace para o efeito do tratamento 1
#   eqm_ale1_var: EQM do método de Laplace para a variância do efeito aleatório
#   eqm_ale2_b0:  EQM do método ML para o intercepto
#   eqm_ale2_b:    EQM do método ML para o efeito do tratamento 1
```

```
#   eqm_ale2_var: EQM do método ML para a variância do efeito aleatório
#   eqm_ale3_b0:  EQM do método REML para o intercepto
#   eqm_ale3_b:   EQM do método REML para o efeito do tratamento 1
#   eqm_ale3_var: EQM do método REML para a variância do efeito aleatório

#   percent_varzero: percentual de simulações em que a variância do efeito aleatório foi
                     estimada em zero no modelo misto por Laplace
#   resultados:          arquivo de resultados de um cenário

# -----------------------------------------------------


# ================================================
# Definição de um novo cenário.
# ================================================


# Entrada de dados

rm(list=ls(all=TRUE))

rfixomil = NULL
rale1mil = NULL
rale2mil = NULL
rale3mil = NULL

# Do experimento
ntrat =  2
nrep =   4
nobs =   50
nsimul = 1000

# Do modelo logístico
beta0 =    0.41
beta =     1.59
varEfAle = 2.00

# -----------------------------------------------------

# Parte I: Geração dos dados

nparcela = ntrat * nrep
tObs =       nparcela * nobs

parcela = sort(rep(seq(1, nparcela), nobs))
x =          rep(sort(rep(c(0, 1), nobs)), nrep)
rep =        sort(rep(seq(1, nrep), nobs*ntrat))


# ================================================
# Fazendo 100 simulações para um mesmo cenário.
# ================================================
```

```
rfixo = NULL
rale1 = NULL
rale2 = NULL
rale3 = NULL

for (s in 1:100)
{
 rm("cor_ale1", "cor_ale2", "cor_ale3", "dados", "efAle", "efeAle_est",
      "eta", "fit", "fit1", "fit2", "fit3", "i", "int_ale1","int_ale2",
      "int_ale3", "intcp", "j", "k", "p", "pi", "result_ale1",
      "result_ale2", "result_ale3", "result_fixo", "u", "var_ale1",
      "var_ale2", "var_ale3", "x_ale1", "x_ale2", "x_ale3", "y")

 # inicialização das matrizes
 efAle = matrix(0, nsimul, tObs)
 intcp =  matrix(0 ,nsimul, tObs)
 eta =    matrix(0 ,nsimul, tObs)
 pi =     matrix(0 ,nsimul, tObs)
 k =      matrix(0 ,nsimul, tObs)
 y =      matrix(0 ,nsimul, tObs)

 # simulação das observações
 for (i in 1:nsimul)
 {
  u = rnorm(nparcela, 0, sqrt(varEfAle))     # geração dos efeitos aleatórios nas parcelas
  for (j in 1:tObs)
  {
   p = (j - 1) %/% nobs + 1
   efAle[i,j] = u[p]
  }
 intcp[i,]= beta0 + efAle[i,]
 eta[i,]=    intcp[i,] + beta * x
  pi[i,]=    exp(eta[i,]) / (1 + exp(eta[i,]))
  k[i,]=     runif(tObs)
  y[i,]=     ifelse(k[i,] < pi[i,], 1, 0)
 }

 # estruturação do conjunto de dados gerados
 dados=data.frame(cbind(sort(rep(seq(1, nsimul), tObs)),
                        rep(parcela, nsimul),
                        rep(rep, nsimul),
                        rep(x, nsimul),
                        matrix(t(y), nsimul*tObs, 1)))
 colnames(dados) = c("sim", "parcela", "rep",  "x", "y")

 # ----------------------------------------------------

 library(lme4)

 # Parte II: Ajuste de modelos: modelo de efeito fixo, modelo de efeito aleatório

   #-------------------------------------------------
```

```
  # A) Modelo de efeito fixo

  fit= by(dados, dados$sim, function(x) glm(y ~ x , family = binomial, data = x))

  result_fixo = t(sapply(fit, coef))
  colnames(result_fixo) = c("int_est", "x_est")

  #---------------------------------------------------

  # B) Modelo misto: por Laplace

  fit1 = by(dados, dados$sim, function(x) lmer(y ~ x + ( 1 | parcela),
            family = binomial, data = x ))

  int_ale1 =    matrix(0, nsimul, 1)
  cor_ale1 =    matrix(0, nsimul, 1)
  x_ale1 =      matrix(0, nsimul, 1)
  efeAle_est = matrix(0, nsimul, nparcela)
  for (i in 1:nsimul)
  {
   efeAle_est[i,] = as.matrix(unlist(sapply(fit1, ranef)[i]))
   int_ale1[i] =     mean(as.matrix(unlist(sapply(fit1, coef)[i]))[1:nparcela] –
                          efeAle_est[i,])
   cor_ale1[i] =    cor(unique(efAle[i,]), efeAle_est[i,])
   x_ale1[i] =      round(as.matrix(unlist(sapply(fit1, coef)[i]))[nparcela+1], 4)
  }
  var_ale1 = as.matrix(unlist(sapply(fit1, VarCorr)))

  result_ale1 = cbind(int_ale1, x_ale1, cor_ale1, var_ale1)
  colnames(result_ale1) =  c("int_est", "x_est", "cor", "var")
  rownames(result_ale1) = NULL

  #---------------------------------------------------

detach("package:lme4")
library(nlme)
#nlmeControl(maxIter=200, pnlsMaxIter=100, msMaxIter=100, tolerance=1e-3)

  # C) Modelo misto: por MV

  fit2 = by(dados, dados$sim, function(x) nlme(y ~ (exp(a + b*x) / (1 + exp(a + b*x))),
            data = x, fixed = list(a ~ 1, b ~ 1), random = a ~ 1 | parcela,
            start = list(fixed = c(a = beta0, b = beta)), method = "ML"))

  int_ale2 =    matrix(0, nsimul, 1)
  cor_ale2 =    matrix(0, nsimul, 1)
  x_ale2 =      matrix(0, nsimul, 1)
  efeAle_est = matrix(0, nsimul, nparcela)
  for (i in 1:nsimul)
  {
   efeAle_est[i,] = as.matrix(unlist(lapply(fit2, ranef)[i]))
```

```
  int_ale2[i] =     mean(as.matrix(unlist(lapply(fit2, coef)[i]))[1:nparcela] -
                          efeAle_est[i,])
  cor_ale2[i] =    cor(unique(efAle[i,]), efeAle_est[i,])
  x_ale2[i] =      round(as.matrix(unlist(lapply(fit2, coef)[i]))[nparcela + 1], 4)
 }
 var_ale2 = as.matrix(as.numeric(t(unlist(sapply(fit2, VarCorr)))[,1]), nsimul, 1)

 result_ale2 = cbind(int_ale2, x_ale2, cor_ale2, var_ale2)
 colnames(result_ale2) = c("int_est", "x_est", "cor", "var")

 #---------------------------------------------------

 # D) Modelo misto: MV restrita

 fit3 = by(dados, dados$sim, function(x) nlme(y ~ (exp(a + b*x) / (1 + exp(a + b*x))),
           data = x, fixed = list(a ~ 1, b ~ 1), random = a ~ 1 | parcela,
           start = list(fixed = c(a = beta0, b = beta)), method = "REML"))

 int_ale3 =    matrix(0, nsimul, 1)
 cor_ale3 =    matrix(0, nsimul, 1)
 x_ale3 =      matrix(0, nsimul, 1)
 efeAle_est = matrix(0, nsimul, nparcela)
 for (i in 1:nsimul)
 {
  efeAle_est[i,] = as.matrix(unlist(lapply(fit3, ranef)[i]))
  int_ale3[i] =     mean(as.matrix(unlist(lapply(fit3, coef)[i]))[1:nparcela] -
                          efeAle_est[i,])
  cor_ale3[i] =    cor(unique(efAle[i,]), efeAle_est[i,])
  x_ale3[i] =      round(as.matrix(unlist((lapply(fit3, coef)[i])))[nparcela + 1], 4)
 }
 var_ale3 = as.matrix(as.numeric(t(unlist(sapply(fit3, VarCorr)))[,1]), nsimul, 1)

 result_ale3 = cbind(int_ale3, x_ale3, cor_ale3, var_ale3)
 colnames(result_ale3) = c("int_est", "x_est", "cor", "var")

 #---------------------------------------------------

 detach("package:nlme")

 rfixo = rbind(rfixo, result_fixo)
 rale1 = rbind(rale1, result_ale1)
 rale2 = rbind(rale2, result_ale2)
 rale3 = rbind(rale3, result_ale3)
}


# ============================================
# Fim das cem simulações.
# Verificação de erros para concatenação de dados.
# ============================================
```

```
warnings()
dim(rfixo)
dim(rale1)
dim(rale2)
dim(rale3)


# ================================================
# Concatenação para armazenamento de dados, caso não haja erro de convergência.
# ================================================


  rfixomil = rbind(rfixomil, rfixo)
  rale1mil = rbind(rale1mil, rale1)
  rale2mil = rbind(rale2mil, rale2)
  rale3mil = rbind(rale3mil, rale3)

  dim(rfixomil)                          # verificação da dimensão do arquivo de dados


# ================================================
# Fim do processo de simulação, análise e armazenamento de dados de um único
    cenário.
# ================================================


# ------------------------------------------------------

# Parte III: Cálculo de médias e EQM

média_fixo = colSums(rfixomil)/1000
média_ale1 = colSums(rale1mil)/1000
média_ale2 = colSums(rale2mil)/1000
média_ale3 = colSums(rale3mil)/1000

eqm_fixo_b0  = sum((rfixomil[1:1000,1]-beta0)^2)/1000
eqm_fixo_b   =  sum((rfixomil[1:1000,2]-beta)^2)/1000
eqm_ale1_b0  = sum((rale1mil[1:1000,1]-beta0)^2)/1000
eqm_ale1_b   =  sum((rale1mil[1:1000,2]-beta)^2)/1000
eqm_ale1_var = sum((rale1mil[1:1000,4] - varEfAle)^2)/1000
eqm_ale2_b0  = sum((rale2mil[1:1000,1]-beta0)^2)/1000
eqm_ale2_b   =  sum((rale2mil[1:1000,2]-beta)^2)/1000
eqm_ale2_var = sum((rale2mil[1:1000,4] - varEfAle)^2)/1000
eqm_ale3_b0  = sum((rale31mil[1:1000,1]-beta0)^2)/1000
eqm_ale3_b   =  sum((rale3mil[1:1000,2]-beta)^2)/1000
eqm_ale3_var = sum((rale3mil[1:1000,4] - varEfAle)^2)/1000

percent_varzero = (dim(rale1mil[rale1mil[,3] == "NA",])[1])/1000

resultados = data.frame(rbind(round(cbind(c(média_fixo, NA, NA),
                                          média_ale1, média_ale2, média_ale3), 2),
                              c(eqm_fixo_b0,eqm_ale1_b0, eqm_ale2_b0, eqm_ale3_b0),
```

```
                                  c(eqm_fixo_b,eqm_ale1_b, eqm_ale2_b, eqm_ale3_b),
                                  c(eqm_fixo_var,eqm_ale1_var, eqm_ale2_var,
                                    eqm_ale3_var),
                                  c(NA, percent_varzero,NA, NA)))
rownames(resultados) = c("int_est", "x_est", "correl", "var_est", "eqm b0", "eqm b",
                                 "eqm var", "% varzero")
colnames(resultados) = c("fixo", "ale1", "ale2", "ale3")

# ------------------------------------------------------

# Parte IV: Exportação dos arquivos de resultados

write.table(rfixomil, "result_fixo_c33.xls", quote = FALSE, row.names = FALSE)
write.table(rale1mil, "result_ale1_c33.xls", quote = FALSE, row.names = FALSE)
write.table(rale2mil, "result_ale2_c33.xls", quote = FALSE, row.names = FALSE)
write.table(rale3mil, "result_ale3_c33.xls", quote = FALSE, row.names = FALSE)
write.table(resultados, "resultadosc33.xls", quote = FALSE)

# ------------------------------------------------------
```

APÊNDICE B – Tabelas relativas ao capítulo 4

Tabela B1 – Cenários para simulação

Cenário	β_0	β	σ_u^2	π_0	π_1
1	-2,20	0,20	0,10		
2	-2,20	0,20	1,00		0,12
3	-2,20	0,20	2,00		
4	-2,20	1,58	0,10		
5	-2,20	1,58	1,00		0,35
6	-2,20	1,58	2,00		
7	-2,20	2,20	0,10		
8	-2,20	2,20	1,00	0,10	0,50
9	-2,20	2,20	2,00		
10	-2,20	2,82	0,10		
11	-2,20	2,82	1,00		0,65
12	-2,20	2,82	2,00		
13	-2,20	4,19	0,10		
14	-2,20	4,19	1,00		0,88
15	-2,20	4,19	2,00		
16	-0,85	0,23	0,10		
17	-0,85	0,23	1,00		0,35
18	-0,85	0,23	2,00		
19	-0,85	0,85	0,10		
20	-0,85	0,85	1,00		0,50
21	-0,85	0,85	2,00		
22	-0,85	1,47	0,10	0,30	
23	-0,85	1,47	1,00		0,65
24	-0,85	1,47	2,00		
25	-0,85	2,84	0,10		
26	-0,85	2,84	1,00		0,88
27	-0,85	2,84	2,00		
28	0,41	0,21	0,10		
29	0,41	0,21	1,00		0,65
30	0,41	0,21	2,00		
31	0,41	1,59	0,10	0,60	
32	0,41	1,59	1,00		0,88
33	0,41	1,59	2,00		

Tabela B2 – Média das estimativas e da correlação entre efeito aleatório real e estimado e percentual de variâncias do efeito aleatório estimadas como zero

(continua)

Cenário 1			
β_0 = -2,20 β = 0,20 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,22	-2,18	-2,18
$\hat{\beta}$	0,22	0,22	0,22
Correlação	NA	0,51	0,51
$\hat{\sigma}_u^2$	0,07	0,06	0,06
$\hat{\sigma}_u^2$=0(%)	0,47	NA	NA

Cenário 2			
β_0 = -2,20 β = 0,20 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,20	-2,01	-2,01
$\hat{\beta}$	0,23	0,20	0,20
Correlação	NA	0,80	0,80
$\hat{\sigma}_u^2$	0,68	0,49	0,49
$\hat{\sigma}_u^2$=0(%)	0,03	NA	NA

Cenário 3			
β_0 = -2,20 β = 0,20 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,20	-1,91	-1,91
$\hat{\beta}$	0,17	0,14	0,14
Correlação	NA	0,83	0,83
$\hat{\sigma}_u^2$	1,51	0,98	0,98
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 4			
β_0 = -2,20 β = 1,58 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,22	-2,19	-2,19
$\hat{\beta}$	1,60	1,57	1,57
Correlação	NA	0,54	0,54
$\hat{\sigma}_u^2$	0,06	0,07	0,07
$\hat{\sigma}_u^2$=0(%)	0,37	NA	NA

Cenário 5			
β_0 = -2,20 β = 1,58 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,20	-1,99	-1,99
$\hat{\beta}$	1,62	1,44	1,44
Correlação	NA	0,81	0,81
$\hat{\sigma}_u^2$	0,76	0,59	0,60
$\hat{\sigma}_u^2$=0(%)	0,02	NA	NA

Cenário 6			
β_0 = -2,20 β = 1,58 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,18	-1,90	-1,90
$\hat{\beta}$	1,56	1,35	1,35
Correlação	NA	0,85	0,85
$\hat{\sigma}_u^2$	1,51	1,11	1,11
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 7			
β_0 = -2,20 β = 2,20 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,21	-2,17	-2,17
$\hat{\beta}$	2,2	2,17	2,17
Correlação	NA	0,54	0,54
$\hat{\sigma}_u^2$	0,06	0,07	0,07
$\hat{\sigma}_u^2$=0(%)	0,37	NA	NA

Cenário 8			
β_0 = -2,20 β = 2,20 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,22	-2,02	-2,02
$\hat{\beta}$	2,24	2,03	2,03
Correlação	NA	0,81	0,81
$\hat{\sigma}_u^2$	0,74	0,60	0,60
$\hat{\sigma}_u^2$=0(%)	0,02	NA	NA

Cenário 9			
β_0 = -2,20 β = 2,20 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,17	-1,88	-1,88
$\hat{\beta}$	2,13	1,84	1,84
Correlação	NA	0,85	0,85
$\hat{\sigma}_u^2$	1,47	1,08	1,09
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Tabela B2 – Média das estimativas e da correlação entre efeito aleatório real e estimado e percentual de variâncias do efeito aleatório estimadas como zero

(continuação)

Cenário 10			
β_0 = -2,20 β =2,82 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,21	-2,17	-2,17
$\hat{\beta}$	2,82	2,77	2,77
Correlação	NA	0,53	0,53
$\hat{\sigma}_u^2$	0,07	0,07	0,07
$\hat{\sigma}_u^2$=0(%)	0,36	NA	NA

Cenário 11			
β_0 = -2,20 β =2,82 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,17	-1,97	-1,97
$\hat{\beta}$	2,81	2,56	2,56
Correlação	NA	0,82	0,82
$\hat{\sigma}_u^2$	0,73	0,57	0,58
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 12			
β_0 = -2,20 β =2,82 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,26	-1,94	-1,94
$\hat{\beta}$	2,86	2,48	2,48
Correlação	NA	0,85	0,85
$\hat{\sigma}_u^2$	1,58	1,13	1,13
$\hat{\sigma}_u^2$=0(%)	0,00	NA	NA

Cenário 13			
β_0 = -2,20 β = 4,19 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,23	-2,20	-2,19
$\hat{\beta}$	4,21	4,14	4,14
Correlação	NA	0,51	0,51
$\hat{\sigma}_u^2$	0,06	0,06	0,06
$\hat{\sigma}_u^2$=0(%)	0,47	NA	NA

Cenário 14			
β_0 = -2,20 β = 4,19 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,23	-2,04	-2,04
$\hat{\beta}$	4,23	3,87	3,87
Correlação	NA	0,80	0,80
$\hat{\sigma}_u^2$	0,74	0,52	0,53
$\hat{\sigma}_u^2$=0(%)	0,03	NA	NA

Cenário 15			
β_0 = -2,20 β = 4,19 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-2,18	-1,90	-1,90
$\hat{\beta}$	4,17	3,64	3,64
Correlação	NA	0,84	0,84
$\hat{\sigma}_u^2$	1,46	0,95	0,96
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 16			
β_0 = -0,85 β = 0,23 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,84	-0,83	-0,83
$\hat{\beta}$	0,22	0,21	0,21
Correlação	NA	0,62	0,62
$\hat{\sigma}_u^2$	0,06	0,06	0,06
$\hat{\sigma}_u^2$=0(%)	0,29	NA	NA

Cenário 17			
β_0 = -0,85 β = 0,23 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,84	-0,78	-0,78
$\hat{\beta}$	0,22	0,20	0,20
Correlação	NA	0,86	0,86
$\hat{\sigma}_u^2$	0,74	0,62	0,62
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 18			
β_0 = -0,85 β = 0,23 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,87	-0,76	-0,76
$\hat{\beta}$	0,24	0,21	0,21
Correlação	NA	0,88	0,88
$\hat{\sigma}_u^2$	1,59	1,22	1,23
$\hat{\sigma}_u^2$=0(%)	0,00	NA	NA

Tabela B2 – Média das estimativas e da correlação entre efeito aleatório real e estimado e percentual de variâncias do efeito aleatório estimadas como zero

(continuação)

Cenário 19			
β_0 = -0,85 β = 0,85 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,84	-0,83	-0,83
$\hat{\beta}$	0,84	0,83	0,83
Correlação	NA	0,65	0,65
$\hat{\sigma}_u^2$	0,06	0,06	0,06
$\hat{\sigma}_u^2$=0(%)	0,25	NA	NA

Cenário 20			
β_0 = -0,85 β = 0,85 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,86	-0,79	-0,79
$\hat{\beta}$	0,88	0,81	0,81
Correlação	NA	0,86	0,86
$\hat{\sigma}_u^2$	0,75	0,63	0,64
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 21			
β_0 = -0,85 β = 0,85 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,83	-0,73	-0,73
$\hat{\beta}$	0,83	0,74	0,74
Correlação	0,88	0,88	0,88
$\hat{\sigma}_u^2$	1,51	1,18	1,19
$\hat{\sigma}_u^2$=0(%)	0,00	NA	NA

Cenário 22			
β_0 = -0,85 β= 1,47 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,84	-0,83	-0,83
$\hat{\beta}$	1,46	1,43	1,43
Correlação	NA	0,62	0,62
$\hat{\sigma}_u^2$	0,06	0,05	0,06
$\hat{\sigma}_u^2$=0(%)	0,28	NA	NA

Cenário 23			
β_0 = -0,85 β = 1,47 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,85	-0,78	-0,78
$\hat{\beta}$	1,47	1,36	1,36
Correlação	NA	0,87	0,87
$\hat{\sigma}_u^2$	0,77	0,64	0,64
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 24			
β_0 = -0,85 β = 1,47 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,87	-0,77	-0,77
$\hat{\beta}$	1,54	1,36	1,36
Correlação	NA	0,89	0,89
$\hat{\sigma}_u^2$	1,54	1,18	1,19
$\hat{\sigma}_u^2$=0(%)	0,00	NA	NA

Cenário 25			
β_0 = -0,85 β = 2,84 σ_u^2 = 0,10			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,84	-0,83	-0,83
$\hat{\beta}$	2,84	2,79	2,79
Correlação	NA	0,57	0,57
$\hat{\sigma}_u^2$	0,07	0,07	0,07
$\hat{\sigma}_u^2$=0(%)	0,34	NA	NA

Cenário 26			
β_0 = -0,85 β = 2,84 σ_u^2 = 1,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,84	-0,78	-0,78
$\hat{\beta}$	2,86	2,62	2,62
Correlação	NA	0,83	0,83
$\hat{\sigma}_u^2$	0,72	0,58	0,58
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 27			
β_0 = -0,85 β = 2,84 σ_u^2 = 2,00			
	Laplace	ML	REML
$\hat{\beta}_0$	-0,85	-0,76	-0,76
$\hat{\beta}$	2,89	2,53	2,53
Correlação	NA	0,85	0,85
$\hat{\sigma}_u^2$	1,46	1,06	1,06
$\hat{\sigma}_u^2$=0(%)	0,00	NA	NA

Tabela B2 – Média das estimativas e da correlação entre efeito aleatório real e estimado e percentual de variâncias do efeito aleatório estimadas como zero

(conclusão)

Cenário 28 $\beta_0 = 0{,}41 \quad \beta = 0{,}21 \quad \sigma_u^2 = 0{,}10$			
	Laplace	ML	REML
$\hat{\beta}_0$	0,42	0,41	0,41
$\hat{\beta}$	0,20	0,20	0,20
Correlação	NA	0,66	0,66
$\hat{\sigma}_u^2$	0,06	0,06	0,06
$\hat{\sigma}_u^2$=0(%)	0,25	NA	NA

Cenário 29 $\beta_0 = 0{,}41 \quad \beta = 0{,}21 \quad \sigma_u^2 = 1{,}00$			
	Laplace	ML	REML
$\hat{\beta}_0$	0,42	0,39	0,39
$\hat{\beta}$	0,19	0,17	0,17
Correlação	NA	0,87	0,87
$\hat{\sigma}_u^2$	0,77	0,66	0,66
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Cenário 30 $\beta_0 = 0{,}41 \quad \beta = 0{,}21 \quad \sigma_u^2 = 2{,}00$			
	Laplace	ML	REML
$\hat{\beta}_0$	0,39	0,34	0,34
$\hat{\beta}$	0,24	0,22	0,22
Correlação	NA	0,89	0,89
$\hat{\sigma}_u^2$	1,57	1,24	1,25
$\hat{\sigma}_u^2$=0(%)	0,00	NA	NA

Cenário 31 $\beta_0 = 0{,}41 \quad \beta = 1{,}59 \quad \sigma_u^2 = 0{,}10$			
	Laplace	ML	REML
$\hat{\beta}_0$	0,41	0,40	0,40
$\hat{\beta}$	1,60	1,57	1,57
Correlação	NA	0,55	0,55
$\hat{\sigma}_u^2$	0,06	0,07	0,07
$\hat{\sigma}_u^2$=0(%)	0,35	NA	NA

Cenário 32 $\beta_0 = 0{,}41 \quad \beta = 1{,}59 \quad \sigma_u^2 = 1{,}00$			
	Laplace	ML	REML
$\hat{\beta}_0$	0,41	0,38	0,38
$\hat{\beta}$	1,62	1,46	1,46
Correlação	NA	0,83	0,83
$\hat{\sigma}_u^2$	0,74	0,59	0,60
$\hat{\sigma}_u^2$=0(%)	0,02	NA	NA

Cenário 33 $\beta_0 = 0{,}41 \quad \beta = 1{,}59 \quad \sigma_u^2 = 2{,}00$			
	Laplace	ML	REML
$\hat{\beta}_0$	0,42	0,38	0,38
$\hat{\beta}$	1,56	1,34	1,34
Correlação	NA	0,86	0,86
$\hat{\sigma}_u^2$	1,54	1,13	1,14
$\hat{\sigma}_u^2$=0(%)	0,01	NA	NA

Tabela B3 – EQM na estimativa do intercepto

Cenário	β_0	Laplace	ML	REML
4	-2,20	0,09	0,09	0,09
5	-2,20	0,35	0,32	0,32
6	-2,20	0,65	0,57	0,57
19	-0,85	0,05	0,05	0,05
20	-0,85	0,31	0,26	0,26
21	-0,85	0,55	0,43	0,43
31	0,41	0,05	0,04	0,04
32	0,41	0,29	0,25	0,25
33	0,41	0,58	0,46	0,46

Tabela B4 – EQM na estimativa do parâmetro regressor

Cenário	β	Laplace	ML	REML
4	1,58	0,14	0,13	0,13
5	1,58	0,59	0,50	0,50
6	1,58	1,12	0,90	0,90
19	0,85	0,10	0,09	0,09
20	0,85	0,59	0,49	0,49
21	0,85	1,15	0,91	0,91
31	1,59	0,13	0,12	0,12
32	1,59	0,65	0,55	0,55
33	1,59	1,17	0,92	0,92

Tabela B5 – EQM na estimativa da variância do efeito aleatório

Cenário	σ_u^2	Laplace	ML	REML
4	0,10	0,01	0,01	0,01
5	1,00	0,43	0,35	0,35
6	2,00	1,62	1,35	1,34
19	0,10	0,01	0,01	0,01
20	1,00	0,34	0,30	0,30
21	2,00	1,24	1,15	1,15
31	0,10	0,01	0,01	0,01
32	1,00	0,36	0,33	0,33
33	2,00	1,55	1,29	1,28