Kiese da S. Quiavauca

Modelo de Regressão Logística com erro de classificação na variável resposta

Niterói - RJ, Brasil

05 de agosto de 2013

Universidade Federal Fluminense

Kiese da S. Quiavauca

Modelo de Regressão Logística com erro de classificação na variável resposta

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientadora: Profa. Ludmilla Jacobson

Coorientador: Prof. Leonardo Bastos

Niterói - RJ, Brasil

05 de agosto de 2013

Universidade Federal Fluminense

Kiese da S. Quiavauca

Modelo de Regressão Logística com erro de classificação na variável resposta

Monografia de Projeto Final de Graduação sob o título *"Modelo de Regressão Logística com erro de classificação na variável resposta"*, defendida por Kiese da S. Quiavauca e aprovada em 05 de agosto de 2013, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Ludmilla Jacobson
Orientadora
Departamento de Estatística – UFF

Prof. Phd. Leonardo Bastos
Coorientador
ENSP – FioCruz

Profa. Dra. Jéssica Kubrusly
Departamento de Estatística – UFF

Prof. Dr. Luis Guillermo Coca Velarde
Departamento de Estatística – UFF

Niterói, 05 de agosto de 2013

Quiavauca, Kiese da Silva
Modelo de Regressão Logística com erro de classificação na variável resposta / Kiese da Silva Quiavauca; Ludmilla da Silva Viana Jacobson, orientadora; Leonardo Soares Bastos, coorientador. Niterói, 2013.
70 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2013.

1. Regressão logística. 2. Sensibilidade. 3. Especificidade. I. Jacobson, Ludmilla da Silva Viana, orientadora. II. Bastos, Leonardo Soares, coorientador. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD -

Resumo

Em grande parte dos estudos relacionados a área da saúde, a ferramenta mais utilizada para detectar se um indivíduo tem ou não o desfecho de interesse é o teste de diagnóstico. Este por sua vez, nem sempre gera dados absolutamente confiáveis. Apesar disso a incerteza presente nos dados gerados por tais testes é conhecida através da sensibilidade (probabilidade do teste dar positivo dado que o indivíduo de fato apresenta o desfecho) e da especificidade (probabilidade do teste dar negativo dado que indivíduo de fato não apresenta o desfecho) do teste utilizado. O objetivo do presente trabalho é acrescentar a incerteza "conhecida" através da sensibilidade e da especificidade no modelo de regressão logística (modelo mais utilizado quando a variável resposta é binária), comparando o modelo que propõe uma correção nas estimativas de prevalência e de razão de chances dos estudos que utilizam testes diagnóstico com o modelo de regressão logística usual. Através dessa comparação foi possível observar que dependendo dos valores de sensibilidade e especificidade do teste utilizado é mais seguro utilizar o modelo corrigido para fazer as análises.

Palavras-chaves: Regressão logística, sensibilidade, especificidade.

Dedicatória

Dedico este trabalho a minha mãe que acreditou em mim, e na minha capacidade antes mesmo que eu, ou outra pessoa pudesse enxergar o que estaria por vir. E ao meu pai, que tem sido exemplo de caráter, força, perseverança e superação não só agora, mas durante toda a minha vida.

Agradecimentos

Agradeço primeiramente a Deus por me permitir viver momentos como esse, e por ter sido o meu sustentador a cada momento.

Agradeço a vó Jod, tia Nusa, tia Jane, tio Marcelo e tia Mirtes que por vezes abriram mão das suas tarefas para me possibilitar terminar esse trabalho.

Agradeço a minha mãe e ao meu padastro Juan que sempre me deram suporte na vida e no estudo, não só durante a minha graduação, mas também durante toda a minha formação no ensino fundamental e médio.

Agradeço a minha irmã Jéssica, que mesmo com sua pouca idade conseguiu me apoiar respeitando os meus momentos de estudo, além de estar sempre disposta a me ajudar quando era preciso.

Agradeço ao meu futuro esposo Júlio César, que tem sido amigo, companheiro e apoio durante estes 6 anos em que estamos juntos.

Agradeço á todos os professores que me deram aula na graduação. Em especial, aos professores: Ludmilla, Leo e Adrian com os quais eu pude aprender muito não só sobre a estatística, mas também sobre a vida.

Agradeço aos meus amigos de caminhada, Carolina Valani, Danielle Freitas, Evandro Dalbem, Guilherme Souza e Marcela Martins que choraram e sorriram comigo durante a minha estadia na UFF. Tenho certeza que a nossa amizade não ficará só na lembrança.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A utilização do modelo de regressão logística em estudos da área da saúde tem sido cada vez mais frequente. A fácil interpretação dos parâmetros estimados através da razão de chances é uma das razões pela qual esse modelo vem sendo escolhido, até para casos onde ele não seria exatamente o mais apropriado. Com intuito de facilitar a interpretação dos resultados, alguns pesquisadores dicotomizam a sua variável resposta para poder então utilizar esse modelo estatístico.

O modelo de regressão logística também é utilizado em estudos em que a variável resposta é resultado de testes de diagnóstico.

Na maior parte das vezes, os testes de diagnóstico utilizados para gerar as observações da variável resposta do modelo de regressão logística não são aqueles considerados os melhores possíveis para detectar a presença ou não da variável do desfecho de saúde de interesse. Por esse motivo, nem sempre tem-se em mãos dados absolutamente confiáveis.

Realizar o ajuste do modelo de regressão logística considerando como variável resposta o resultado de testes desse tipo, pode ser perigoso, uma vez que os resultados e as interpretações serão baseadas no testes alternativo, enquanto a intenção é que estas sejam baseadas na presença ou não da variável que realmente interessa. Muitos estudos, no entanto, fazem o uso de testes diagnósticos e não consideram o erro de classificação cometido.

A incerteza presente nos dados gerados por testes alternativos pode ser conhecida através da sensibilidade (probabilidade do teste dar positivo dado que o indivíduo de fato apresenta o desfecho) e da especificidade (probabilidade do teste dar negativo dado que o indivíduo de fato não apresenta o desfecho) do teste utilizado. Essa incerteza conhecida pode ser acrescentada no modelo de regressão logística, o que corrige os erros de classificação cometidos na variável resposta, quando se utiliza testes de diagnóstico **?**.

Neste trabalho será apresentado o modelo de regressão logística corrigido pela sensibilidade e especificidade. Além disso serão feitas comparações dos resultados encontrados

através da utilização dos modelos corrigido e usual.

Este trabalho esta dividido em 7 capítulos. O primeiro corresponde a esta introdução. No capítulo 2 são apresentados os objetivos geral e específicos. O capítulo 3 descreve os métodos estatísticos utilizados nos Modelos Linerares Generalizados, família de modelos a qual o modelo de regressão logística pertence, visando facilitar o entendimento não só do modelo de regressão logística como também do modelo de regressão logística corrigido. Estes modelos serão descritos no captulo 4 com mais detalhes, apresentando também a aplicação dos métodos de MLG em cada um deles. Após a construção da teoria necessária, será apresentado, no capítulo 5, um estudo simulado realizado com o objetivo de melhor entender as diferenças entre os modelos de regressão logística corrigido e usual. Por fim, no capítulo 6, serão expostos os resultados encontrados na aplicação dos modelos corrigido e usual em um banco de dados referente a crianças asmáticas,seguido da conclusão do trabalho, no capítulo 7.

2 Objetivo

2.1 Objetivo Geral

Avaliar um método de correção para o modelo de regressão logística quando a variável resposta se refere a resultados de testes de diagnóstico.

2.2 Objetivos Específicos

Aplicar a metodologia proposta por *Ardo van den Houta et.al (2007)* no modelo de regressão logística ajustando o método para o caso em que a variável resposta é resultado de testes de diagnóstico, utilizando como correção a sensibilidade e especificidade do teste;

Realizar um estudo simulado, para explorar as vantagens/desvantagens do modelo corrigido;

Comparar os resultados obtidos através dos modelos de regressão logística usual e corrigido, a partir de uma aplicação com dados de asma.

3 Modelos Lineares Generalizados

Modelos estatísticos são utilizados com o objetivo de reproduzir de forma simplificada a essência e o comportamento de sistemas complexos alvos do nosso estudo. Fazendo o uso de métodos matemáticos esses modelos têm o propósito de analisar, descrever, explicar e simular uma determinada variável (variável resposta) através do conhecimento prévio de outras variáveis que influenciam a variável de interesse (variáveis explicativas ou covariáveis). Os métodos de modelagem estatística podem ser aplicados em diferentes áreas e situações, por esse motivo se faz cada vez mais frequente a utilização dos mesmos. Durante muitos anos o estudo da relação entre as variáveis explicativas e a variável reposta foi feito principalmente através do modelo de regressão linear, no entanto esse modelo possui algumas restrições que acabam tornando a simplificação da realidade menos verossímil do que o desejado. O modelo de regressão linear segue a seguinte forma *(Gujarati, D.N;2011)*

$$y = X\beta + \varepsilon;$$

onde y é o vetor da variável resposta (dimensão n x 1), X é a matriz de covariáveis (dimensão n x p), β é o vetor dos coeficientes (dimensão p x 1), n é o número de observações e ε é o erro aleatório que segue a distribuição normal multivariada com média $\underline{0}$ e matriz de covariâncias $\sum$. Para que esse modelo tenha validade é necessário também que a hipótese de independência dos erros aleatórios (entre si e com o vetor de variáveis explicativas) seja satisfeita. Em geral, se a condição de normalidade não for satisfeita procura-se uma transformação que possibilite que isso aconteça. No entanto, nem sempre é tão simples encontrar esta transformação. Por isso, foram criados modelos estatísticos mais específicos com outras suposições e hipóteses. A união de alguns desses modelos resultou no conjunto dos modelos lineares generalizados, que tem como caso particular o modelo de regressão linear.

3.1 Família Exponencial

3.1.1 Definição

A família exponencial se trata de um conjunto de distribuições com algumas características em comum.

Definição *(Dobson, A.J; 2002) :* Seja Y uma variável aleatória cuja $f.d.p$ (função de probabilidade, se Y é discreta ou função de densidade, se Y é contínua) depende do parâmetro θ. Dizemos que a $f.d.p$ de Y pertence à família exponencial se pudermos escrevê-la da seguinte forma:

$$f(y,\theta) = exp\{a(y)b(\theta) + c(\theta) + d(y)\}, \tag{3.1}$$

onde $a(.)$, $b(.)$, $c(.)$, $d(.)$, são funções conhecidas. Se a função $a(y) = y$ dizemos que a distribuição está na forma canônica, neste caso, a função $b(\theta)$ é chamada de parâmetro natural.

Algumas das distribuições que pertencem a família exponencial são: normal, poisson, exponencial, binomial e outras. O foco desse trabalho será na distribuição binomial, uma vez que os objetivos se referem ao modelo de regressão logística, cuja a variável resposta é binária.

3.1.2 Esperança e Variância

Podemos encontrar uma forma geral para o valor esperado e a variância de $a(Y)$. Se a distribuição é da forma canônica, por exemplo, conhecer $E[a(Y)]$ e $V[a(Y)]$ significa conhecer $E[Y]$ e $V[Y]$. Segundo *Dobson, A. J (2002)*, para encontrar essas expressões utilizam-se alguns resultados envolvendo função de densidade. Da definição de função de densidade tem-se que

$$\int_{-\infty}^{\infty} f(y,\theta)dy = 1, \tag{3.2}$$

(se a variável for discreta a integral será substituída pelo somatório). Além disso, ao derivar os dois lados de (3.2) com relação a θ observa-se que

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(y,\theta)dy = \frac{d}{d\theta}1 = 0. \tag{3.3}$$

Alterando a ordem da derivada e da integral na primeira parte de (3.3), obtemos

$$\int_{-\infty}^{\infty} \frac{df(y,\theta)}{d\theta} dy = 0. \tag{3.4}$$

De forma similar, ao derivar (3.2) duas vezes em relação a θ e inverter a ordem da derivada e da integral, tem-se

$$\int_{-\infty}^{\infty} \frac{d^2 f(y,\theta)}{d\theta^2} dy = 0. \tag{3.5}$$

Em especial, por se tratar da familia exponencial,

$$f(y,\theta) = exp\{a(y)b(\theta) + c(\theta) + d(y)\}, \tag{3.6}$$

e consequentemente

$$\begin{aligned} \frac{d}{d\theta} f(y,\theta) &= exp\{a(y)b(\theta) + c(\theta) + d(y)\}(a(y)b'(\theta) + c'(\theta)) & (3.7)\\ &= f(y,\theta)(a(y)b'(\theta) + c'(\theta)), & (3.8) \end{aligned}$$

De (3.4), tem-se que:

$$\int_{-\infty}^{\infty} f(y,\theta)(a(y)b'(\theta) + c'(\theta))dy = 0.$$

Basta notar que

$$\int_{-\infty}^{\infty} f(y,\theta)(a(y)b'(\theta) + c'(\theta))dy = 0$$

$$\Longrightarrow \int_{-\infty}^{\infty} f(y,\theta)a(y)b'(\theta)dy + \int_{-\infty}^{\infty} f(y,\theta)c'(\theta)dy = 0$$

$$\Longrightarrow b'(\theta)\int_{-\infty}^{\infty} f(y,\theta)a(y)dy + c'(\theta)\int_{-\infty}^{\infty} f(y,\theta)dy = 0$$

$$\Longrightarrow b'(\theta)E[a(Y)] + c'(\theta) = 0.$$

Com isso tem-se que $E[a(Y)] = -c'(\theta)/b'(\theta)$. Utiliza-se argumentos similares para obter

a expressão de $V[a(Y)]$. Note que

$$\frac{d^2 f(y,\theta)}{d\theta^2} = (a(y)b''(\theta) + c''(\theta))f(y,\theta) + (a(y)b'(\theta) + c'(\theta))^2 f(y,\theta).$$

Além disso, $(a(y)b'(\theta) + c'(\theta))^2 f(y,\theta) = (b'(\theta)^2(a(y) - E[a(Y)])^2 f(y,\theta)$. Assim sendo, de (3.5), da expressão de $E[a(Y)]$ e da definição de variância temos que

$$\int_{-\infty}^{\infty} \frac{d^2 f(y,\theta)}{d\theta^2} dy = b''(\theta)E[a(Y)] + c''(\theta) + (b'(\theta))^2 V[a(Y)].$$

O que implica

$$V[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)^3}.$$

3.1.3 Função Score

Partindo da definição de família exponencial, a função de log-verossimilhança desse tipo de distribuições é dada por:

$$l(\theta, y) = a(y)b(\theta) + c(\theta) + d(y). \tag{3.9}$$

A partir da função de log-verossimilhança surge uma nova função que será utilizada posteriormente na estimação dos parâmetros em modelos lineares generalizados, essa função é denominada função *score*. Tal função é definida da seguinte forma:

$$U(\theta, y) = \frac{dl(\theta, y)}{d\theta} = a(y)b'(\theta) + c'(\theta). \tag{3.10}$$

A função score U é uma variável aleatória, pois é função de Y. Neste caso, é necessário conhecer o seu valor esperado e a sua variância.

$$\begin{aligned} E[U] &= E[a(y)b'(\theta) + c'(\theta)] \\ &= b'(\theta)E[a(y)] + c'(\theta) \\ &= b'(\theta)\left(\frac{-c'(\theta)}{b'(\theta)}\right) + c'(\theta) \\ &= -c'(\theta) + c'(\theta) = 0. \end{aligned}$$

A variância de U, chamada de informação, também pode ser encontrada em função da

variância de a(Y).

$$\begin{aligned} V[U] &= V[a(y)b'(\theta) + c'(\theta)] \\ &= (b'(\theta))^2 V[a(y)] \\ &= \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta). \end{aligned}$$

Outra propriedade de U que será usada posteriormente é a que nos garante que $V[U] = E[U']$, essa igualdade pode ser demonstrada através da definição de variância ($V[X] = E[X^2] - E^2[X]$).

3.1.4 Aplicação na Distribuição Binomial

Sejam Y o número de "sucessos" em n tentativas independentes, e π a probabilidade de sucesso em cada tentativa, podemos dizer que a $f.d.p$ de Y pertence à família exponencial.

Com efeito, basta notar que dado $y = 1, 2, 3, ..., n$

$$\begin{aligned} f(y, \pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \exp\left\{ \log\binom{n}{y} + y\log(\pi) + (n-y)\log(1-\pi) \right\} \\ &= \exp\left\{ y\log(\pi) - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y} \right\} \\ &= \exp\left\{ y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y} \right\}, \end{aligned}$$

obtendo-se,
$a(y) = y$, $b(\pi) = \log(\pi/(1-\pi))$, $c(\pi) = n\log(1-\pi)$ e $d(y) = \log\binom{n}{y}$. É possível notar que a distribuição de Y é da forma canônica, e que $\log(\pi/(1-\pi))$ é o parâmetro natural desta distribuição.

Além disso, como $b'(\pi) = \dfrac{1}{\pi(1-\pi)}$, $c'(\pi) = \dfrac{-n}{1-\pi}$, $b''(\pi) = \dfrac{2\pi - 1}{\pi^2(1-\pi)^2}$ e $c''(\pi) = \dfrac{-n}{(1-\pi)^2}$ tem-se que

$$\begin{aligned} E[a(Y)] &= \frac{-c'(\pi)}{b'(\pi)} \\ &= \frac{-n}{(1-\pi)} \setminus \frac{1}{\pi(1-\pi)} \\ &= n\pi \\ &= E[Y], \end{aligned}$$

e

$$\begin{aligned} V[a(Y)] &= \frac{b''(\pi)c'(\pi) - c''(\pi)b'(\pi)}{b'(\pi)^3} \\ &= \frac{n(1-\pi)}{\pi^2(1-\pi)^3} \setminus \frac{1}{\pi^3(1-\pi)^3} \\ &= n\pi(1-\pi) \\ &= V[Y]. \end{aligned}$$

Como a distribuição binomial está na forma canônica da família exponencial,espera-se que $E[a(Y)] = E[Y]$ e $V[a(Y)] = V[Y]$. Através dos cálculos acima é possível verificar que de fato a esperança e a variância encontradas através da fórmula genérica de esperançaa e variância de distribuições pertencentes á familia exponencial coincide com a esperança e variância de Y , tal que $Y \sim Binomial(n; \pi)$.

3.2 Modelos Lineares Generalizados

3.2.1 Definição

Definição *(Demetrio, C.G.B. (2002); Dobson, A.J (2002)):* Seja Y uma variável aleatória associada a um conjunto de variáveis explicativas $x_1, x_2, x_3, \ldots, x_p$. Para uma amostra de n observações (x_i, y_i), em que $x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})^T$ é o vetor coluna das variáveis explicativas, o modelo linear generalizado se baseia na relação de três componentes:

i. Componente aleatório: representado pelo conjunto de variáveis aleatórias independentes $Y_1, Y_2, Y_3, \ldots, Y_n$ obtidas de uma mesma distribuição que faz parte da família

exponencial e está na forma canônica, com $E[Y_i] = \mu_i, i = 1, 2, 3, .., n$, isto é

$$f(y_i, \theta_i) = exp\left\{y_i b(\theta_i) + c(\theta_i) + d(y_i)\right\}.$$

ii. Componente sistemático: representado através da combinação linear entre as variáveis explanatórias e seus efeitos, $\eta_i = x_i^T \beta$, onde $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)^T$.

iii. Função de ligação: uma função monótona e diferenciável que relaciona o componente aleatório ao componente sistemático, obtendo a seguinte relação $\eta_i = g(\mu_i)$.

É possível observar já na definição de modelos lineares generalizados (MLG) uma importante característica que possibilita utilizar esses modelos em diferentes circunstâncias. A relação entre as componentes é determinada através da função de ligação escolhida, podendo esta ser qualquer função monótona e diferenciável, o que possibilita maior proximidade do modelo com a realidade do estudo em questão. Como caso particular no modelo de regressão linear a função de ligação utilizada é a identidade, isto é, $\eta_i = \mu_i$.

Outro atributo importante e facilitador nos MLG é que neles define-se uma distribuição para a variável resposta que representa as observações e não uma distribuição para o erro aleatório.

Conclui-se da definição de MLG que para construir um modelo desse tipo, basta especificar a distribuição da variável resposta, a matriz do modelo (variáveis explicativas) e a relação entre a variável resposta e as variáveis explicativas. O exemplo a seguir *(Dobson, A.J; 2002)* visa ilustrar o passo a passo necessário para a utilização desses modelos.

Considere uma língua que foi "criada" a partir de uma outra língua já existente (ex. grego moderno e grego antigo). Um modelo simples para a mudança ocorrida no vocabulário é que se a distância entre a língua já existente e a criação da nova língua ocorreu num tempo t então a probabilidade de que existam palavras cognatas entre essas línguas para um particular significado é $e^{-\theta t}$, onde θ é o parâmetro, suponhamos que θ é aproximadamente igual para os significados de uso geral. Para uma amostra de tamanho N com diferentes palavras usualmente utilizadas tem-se uma espécie de juiz que definirá para cada palavra o seu correspondente nas duas línguas e dirá se essas são cognatas ou não.

Neste caso poderíamos dizer que a componente aleatória segue a distribuição Bernoulli ($Y_i = 1$, se as línguas têm cognatas para i-ésimo significado; e $Y_i = 0$, se as palavras não forem cognatas). Além disso, de acordo com a construção do exemplo teríamos

$\pi_i = P(Y_i = 1) = e^{-\theta t}$. Note que desta forma a função de ligação já está praticamente definida, já que t é a variável explicativa e na distribuição Bernoulli $\mu_i = \pi_i$. Sendo assim, a função de ligação pode ser $g(\pi) = log(\pi) = -\theta t$. É importante destacar que essa função de ligação foi escolhida por causa da relação apresentada no exemplo, mas outras funções de ligação poderiam ser escolhidas, para descrever a relação entre o tempo entre a língua antiga e a criação da nova língua e o fato das línguas terem cognatas ou não.

3.2.2 Estimação

A estimação dos parâmetros envolvidos no modelo linear generalizado é feita através do método de máxima verossimilhança. Apesar de ser possível, em alguns casos, encontrar expressões explícitas para os estimadores, normalmente essas expressões são obtidas através do uso de métodos iterativos com base no algorítimo de Newton-Raphson. Segundo *Dobson, A.J (2002)* a estimação dos parâmetros é feita da seguinte forma:

Considerando que amostra aleatória $Y_1, Y_2, ..., Y_n$ satisfaz a definição de modelos lineares generalizados, o objetivo é estimar o vetor de parâmetros β que se relaciona com os $Y_i's$ através da função de ligação, de forma que se $E[Y_i] = \mu_i$, então $g(\mu_i) = x_i^T\beta$.

A função de log-verossimilhança de $Y_i, i = 1, 2, 3, ..., n$ é dada por:

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i), \tag{3.11}$$

sendo $b(.), c(.)$ e $d(.)$ as mesmas funções de (3.1). Além disso, já foi visto que

$$E[Y_i] = E[a(Y_i)] = -c'(\theta_i)/b'(\theta_i), \tag{3.12}$$

$$V[Y_i)] = V[a(Y_i)] = (b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i))/(b'(\theta_i))^3 \tag{3.13}$$

e

$$g(\mu_i) = x_i^T\beta = \eta_i, \tag{3.14}$$

onde x_i é o vetor composto pelos elementos $x_{ij}, j = 1, 2, 3, ..p$. Pode-se dizer que $E[Y_i] = E[a(Y_i)]$ e que $V[Y_i] = V[a(Y_i)]$, $i = 1, 2, 3, ...n$, pois na definição de modelos lineares generalizados a distribuição dos "$Y_i's$" está na forma canônica da família exponencial. A

função de log-verossimilhança conjunta de $Y_1, Y_2, Y_3, ..., Y_n$ é dada por

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i).$$

O objetivo do método da máxima verossimilhança é encontrar o valor de β que maximiza a função de log-verossimilhança dada uma amostra $y_1, y_2, ..., y_n$. Para isso, basta igualar a função score a zero. Note que,

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} . \frac{\partial \theta_i}{\partial \mu_i} . \frac{\partial \mu_i}{\partial \beta_j}, \tag{3.15}$$

sendo que a última igualdade de (3.15) decorre da regra da cadeia. Pensemos agora nos termos que apareceram na última igualdade de (3.15). Primeiro, vejamos que

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i),$$

sendo que a segunda afirmação decorre de (3.12). Continuando, substituindo (3.12) na derivada de (3.11), obtemos

$$\frac{\partial \theta_i}{\partial \mu_i} = 1 / \frac{\partial \mu_i}{\partial \theta_i},$$

derivando (3.12) temos

$$\begin{aligned} \frac{\partial \mu_i}{\partial \theta_i} &= \frac{c''(\theta_i}{b'(\theta_i)} + \frac{c'(\theta_i) b''(\theta_i)}{(b'(\theta_1))^2} \\ &= b'(\theta_i) V[Y_i] \end{aligned}$$

Por fim, de (3.14), temos que

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} . \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} . x_{ij}.$$

Sendo assim, a função score (3.15) pode ser escrita da seguinte forma:

$$U_j = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{V[Y_i]} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]. \tag{3.16}$$

A matriz de variância e covariância dos $U_j's$, também conhecida como matriz de in-

formação, será dada por

$$\begin{aligned} I_{jk} &= E[U_j U_k] \\ &= E\Bigg[\sum_{i=1}^{n}\left[\frac{(Y_i-\mu_i)}{V[Y_i]}x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right]\sum_{i=1}^{n}\left[\frac{(Y_i-\mu_i)}{V[Y_i]}x_{ik}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right]\Bigg] \\ &= \sum_{i=1}^{n}\frac{E((Y_i-\mu_i)^2)x_{ij}x_{ik}}{(V[Y_i])^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2, \end{aligned}$$

com efeito, pois $E[(Y_i-\mu_i)(Y_l-\mu_l)]=0$ para $i\neq l$, já que as Y são independentes. Usando $E[(Y_i-\mu_i)^2]=V[Y_i]$, a matriz de informação é simplificada para

$$I_{jk}=\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{(V[Y_i])}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{3.17}$$

Fazendo o uso de métodos iterativos para encontrar a solução de $U_j=0$ teremos que a estimativa para o vetor de parâmetros será dada por:

$$b^{(m)}=b^{(m-1)}+[I^{(m-1)}]^{-1}U^{(m-1)}, \tag{3.18}$$

onde $b^{(m)}$ é o vetor de estimativas para $\beta_1,...,\beta_p$ na m-ésima iteração. Na equação (3.18), $[I^{(m-1)}]^{-1}$ é a inversa da matriz de informação, cujos elementos são dados em (3.17) e $U^{(m-1)}$ é o vetor de elementos dados em (3.16) todos avaliados em $b^{(m-1)}$.

É possível mostrar que a equação (3.18) pode ser simplificada pela equação abaixo

$$X^TW^{(m-1)}Xb^{(m)}=X^TW^{(m-1)}z^{(m-1)} \tag{3.19}$$

$$\Longrightarrow b^{(m)}=(X^TW^{(m-1)}X)^{-1}X^TW^{(m-1)}z^{(m-1)}, \tag{3.20}$$

onde
$W^{(m-1)}$ é uma matriz diagonal n x n com $w_{ii}=\dfrac{1}{V[Y_i]}\left(\dfrac{\partial\mu_i}{\partial\eta_i}\right)^2$ e $z_i=\eta_i+(y_i-\mu_i)\left(\dfrac{\partial\eta_i}{\partial\mu_i}\right)$.

A equação matricial (3.19) é válida para qualquer MLG e mostra que a solução das equações de MV equivale a calcular repetidamente uma regressão linear ponderada de uma variável dependente ajustada z sobre a matriz X usando uma função de peso W que se modifica no processo iterativo.As funções de variância e de ligação entram no processo iterativo através de W e z. O método usual para iniciar o processo iterativo é especificar uma estimativa inicial e sucessivamente alterá-la até que a convergência seja obtida.

3.2.3 Inferência

A construção de um modelo estatístico envolve três passos, são esses: (1) especificação do modelo; (2) estimação dos parâmetros e (3)inferência. Os dois primeiros passos foram apresentados nas seções anteriores, nesta seção será estudado o terceiro passo. De forma geral, a inferência estatística visa utilizar dados amostrais em conjunto com técnicas de probabilidade para expandir, quando possível, os conceitos encontrados em uma determinada amostra para a população da qual esta foi selecionada. No que tange aos modelos lineares generalizados a inferência estatística é utilizada com o intuito de validar o modelo ajustado através dos 2 primeiros passos citados. Uma vez estimados os paramêtros do modelo, é necessário saber se a presença da variável associada a esse parâmetro é realmente significativa, e verificar se uma determinada variável perde importância no modelo após a inclusão de outra variável.

Para a realização da inferência sobre o modelo, é necessário conhecer a distribuição dos estimadores. No entanto é muito difícil, no caso de MLG, obter a distribuição exata dos estimadores dos parâmetros. Isso acontece pois nos MLG existem infinitas combinações de função de distribuição e função de ligação. Sendo assim, para encontrar a distribuição exata dos estimadores dos parâmetros seria necessário realizar um estudo caso a caso combinando todas as distribuições e funções de ligação possíveis, o que não seria viável. Por esse motivo, os testes de hipóteses e intervalos de confiança dos MLG fazem o uso da distribuição assintótica dos estimadores dos parâmetros.

Tem-se a seguir alguns dos métodos utilizados para inferir sobre MLG e em quais situações estes podem/devem ser utilizados. A maioria desses métodos utilizam estatísticas provenientes da teoria de máxima verossimilhança. Todos os métodos que serão apresentados partem do pressuposto de que a especificação do modelo foi feita de forma correta, assim como descrito na definição de MLG.

3.2.3.1 Teste da Razão de Verossimilhanças

O principal objetivo deste teste é verificar a real importância da inclusão de uma ou mais variáveis no modelo. Sendo assim, a idéia é comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem as variáveis em questão. A comparação dos valores observados com os valores preditos é baseada no log da verossimilhança, ou na *deviance* dos dois modelos. Se a diferença entre os dois modelos for muito grande, então rejeita-se a hipótese nula de que o modelo com menos parâmetros

possui um ajuste tão bom quanto o modelo com mais parâmetros.

A estatística de teste é dada por: $\Lambda = -2ln(L_s) + 2ln(L_c)$, onde L_s é a verossimilhança do modelo sem a(s) covariável(eis) e L_c é a verossimilhança do modelo com a(s) covarivél(eis). Sob H_0, para grandes amostras Λ tem distribuição χ^2_q, sendo q o número de variáveis a mais no modelo que está sendo testado.

De acordo com o teste de razão de verosimilhanças, rejeita-se H_0 a um nível de significância α se o valor observado da estatística de teste Λ for superior ao quantil de probabilidade $1 - \alpha$ da distribuição χ^2_p.

O uso da estatística de razão de verosimilhanças não envolve grandes dificuldades computacionais, já que, para calcular a estatística Λ basta usar o método iterativo de mínimos quadrados ponderados. A estatística de razão de verossimilhanças é mais utilizada para comparar modelos que estão encaixados, isto é, modelos em que um é submodelo do outro.

3.2.3.2 Teste de Wald

Esse teste é baseado na distribuição normal assintótica de $\widehat{\beta}$ e é uma generalização da estatística t de Student. Na maioria das vezes, ele é o mais usado no caso de hipóteses relativas a um único coeficiente . A vantagem desse teste em relação ao teste da razão de verossimilhanças é que nele não é necessário realizar os cálculos para o modelo com e sem a variável candidata.

A estatística de teste é dada por: $W = (\widehat{\beta} - \beta)^T I(\widehat{\beta} - \beta)$, sob H_0 a estatística W tem distribuição assintoticamente qui-quadrado com p graus de liberdade, sendo p o número de parâmetros no modelo. Se $W > \chi^2_{p,1-\alpha}$, a um nível de significância de $\alpha\%$ rejeita-se a hipótese nula.

3.2.3.3 Intervalos de confiança

O intervalo de confiança dos estimadores dos parâmetros dos MLG podem ser deduzidos através das distribuições assintóticas das estatísticas dos testes de hipótese apresentados acima. Além disso, de *Gauss, M.C. (2007)* tem-se que para grandes amostras $\widehat{\beta} \sim N_p(\beta, I^{-1})$. Através dessa distribuição também é possível construir intervalos de confiança para β.

4 Regressão Logística corrigida pela sensibilidade e especificidade

4.1 Regressão Logística

O modelo de regressão logística é um modelo que faz parte dos modelos lineares generalizados e é utilizado quando a variável reposta é uma variável binária. Diz-se que um modelo linear generalizado é o modelo de regressão logística se a variável resposta segue a distribuição Bernoulli, e a função de ligação utilizada é $\eta_i = g(\pi_i) = log(\pi_i/(1-\pi_i))$ $= x_i^T\beta$, a função de ligação do modelo de regressão logística é chamada de função logit.

4.1.1 Estimação

A estimação dos parâmetros no modelo de regressão logística é feita, assim como no caso geral de MLG, através do processo iterativo de mínimos quadrados ponderados $b^{(m)} = (X^TW^{(m-1)}X)^{-1}X^TW^{(m-1)}z^{(m-1)}$, no entanto no caso do modelo de regressão logística os componentes dessa expressão são mais específicos. Dada a distribuição da variável resposta e a função de ligação utilizada neste modelo, tem-se que $W = diag(\pi_1(1-\pi_1), ..., \pi_n(1-\pi_n))$, $z = (z_1, z_2, ..., z_n)^T$ é a variável dependente modificada ,
$\eta_i + (y_i - \pi_i) \setminus \pi_i(1-\pi_i), i = 1, 2, 3, ..., n$.

4.1.2 Razão de Chances

De forma resumida, a razão de chance expressa quais são as chances de que um evento aconteça se, sob as mesmas condições ele não acontecer. Ou seja, a razão de chances é uma medida de associação e expressa a aproximação do quanto é mais provável (ou improvável) para o resultado estar presente entre aqueles que possuem uma determinada característica do que entre aqueles não a possuem. Por exemplo, se y denota a presença ou ausência

de uma determinada espécie e x denota se a área tem ou não tem floresta, o Odds = 2 indica que a presença daquela espécie é duas vezes mais esperada em áreas com floresta do que em áreas sem floresta. Ou seja, a presença de floresta é muito importante para aumentar a chance de ocorrência daquela espécie. Outro exemplo, que talvez possa ser mais intuitivo, seria a razão de chances de ser atropelado toda vez que se atravessa uma avenida. Mesmo que você atravesse a avenida e não seja atropelado, existia uma chance deste evento ocorrer, essa chance é a "razão de chances" ou *"odds ratio"*.

A razão de chances é calculada utilizando a probabilidade de sucesso da variável resposta, tem-se que $OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$, sendo π_1 a probabilidade de sucesso no primeiro grupo e π_2 a probabilidade de sucesso no segundo grupo. Como no modelo de regressão logística $log(\frac{\pi}{1-\pi} = x^T\beta)$, a razão de chance também pode ser calculada em função dos parâmetros estimados, obtendo-se $OR = exp(\beta)$.

O intervalo de confiança da razão de chance pode ser calculado através do exponencial do intervalo de confiância do parâmetro β.

4.2 Erros de classificação

Quando se trata de variáveis dicotômicas pode-se ter alguns problemas na coleta de dados, como por exemplo, erro de classificação na variável resposta. Segue abaixo alguns casos em que este erro pode acontecer.

CASO 1: A variável resposta é coletada através de uma pergunta em um questionário. No entanto esta pergunta leva em consideração algum assunto extremamente pessoal que pode constranger o entrevistado. Sendo assim, dependendo da forma com que essa pergunta for feita e de quem estiver próximo do entrevistado no momento da entrevista existe uma grande possibilidade de que a resposta dada não seja a correta.

CASO 2: A variável resposta deve ser obtida através de um procedimento muito caro (ou muito demorado). No entanto por falta de orçamento (ou de tempo), utiliza-se um procedimento alternativo que tem um custo menor, mas é suscetível a erros.

Os dois casos descritos acima têm uma característica em comum, é possível considerar no modelo de regressão logística os erros que sabemos estão sendo cometidos.

O problema do primeiro caso pode ser solucionado através da utilização da técnica **RR** (*randomized response*) que consiste em criar uma forma de resposta aleatória de modo que seja possível conhecer e controlar a incerteza envolvida com a variável de interesse

Ardo van den Houta, et. al (2007).

Exemplo (RR): No questionário de uma pesquisa a pergunta que será utilizada como variável resposta é sobre a utilização ou não de drogas ilícitas. Para preservar os entrevistados foi utilizada a técnica RR. Sendo assim, a resposta desta pergunta será dada da seguinte forma: sem que o entrevistador veja, o entrevistado lançará dois dados e fará a soma dos resultados obtidos; se a soma dos resultados for 2,3 ou 4 o entrevistado responderá sim, se a soma dos resultados for 5,6,7,8,9 ou 10 o entrevistado responderá a verdade e por fim, se a soma dos resultados for 11 ou 12 o entrevistado responderá não.

Fazendo o uso desta técnica tem-se o controle dos erros que estarão presentes na variável resposta. Já que utilizando o teorema da probabilidade total será possível escrever $\pi_i = P(Y_i = 1)$ em função da variável D_i, e consequentemente em função da variável Z_i, onde:

Y_i representa a resposta obtida na pergunta (usando o método RR) do i-ésimo entrevistado (1, se sim; 0, se não),

D_i, representa o resultado obtido através da soma dos dois dados

e Z_i representa a resposta que o entrevistado daria se dissesse a verdade.

Tendo essa função em mãos ficará fácil escolher uma função de ligação capaz de captar essa informação já que $\pi_i = \mu_i$ no caso da regressão logística. Desta forma basta escolher as variáveis explicativas e tem-se o modelo completo para fazer as análises desejadas.

No caso 2, pode-se utilizar os conceitos de sensibilidade e especificidade para que a análise seja feita considerando a variável que realmente interessa (ex. o indivíduo tem ou não a doença) e não a variável que tem-se em mãos (ex. o teste (procedimento alternativo) deu ou não positivo). Tendo como foco o segundo caso citado, o presente estudo irá mostrar uma forma de corrigir o possível erro cometido na coleta da variável resposta.

4.3 Regressão Logística corrigida pela sensibilidade e especificidade

4.3.1 Sensibilidade e Especificidade

Em geral estudos que tem por objetivo analisar as variáveis associadas com a ocorrência ou não de determinada doença, precisam se preocupar também com a coleta dos dados responsáveis por nos indicar se o entrevistado/paciente tem ou não a doença em questão.

O ideal é que a coleta desses dados seja feita através do uso do teste com o melhor critério diagnóstico disponível (teste padrão-ouro), no entanto a utilização deste tipo de teste nem sempre é fácil, rápida e de baixo custo, por esse motivo em muitos desses estudos são utilizados testes alternativos na coleta dos dados.

Exemplo 1 (Teste do suor) ***(Gianni Mastella, 2010)***: O teste do suor é o principal teste para diagnóstico da fibrose cística (FC). No entanto este teste que ainda é considerado padrão-ouro tem sido repetidamente criticado por causa de sua complexidade, já que o método inclui a necessidade de extrair suor de filtros de papel ou gaze usados para coletá-lo, pesar as amostras duas vezes antes e após a coleta, além de outros procedimentos. Por essa razão, tentativas têm sido feitas para estabelecer um teste mais simples para medir a concentração de eletrólitos no suor.

Juntamente com os testes alternativos, vêm a incerteza sobre a variável resposta, o fato do teste dar positivo (ou negativo) não nos dá a certeza de que de fato o indivíduo realmente apresente (ou não) a característica de interesse. Por esse motivo ao realizar testes diagnósticos alternativos é importante mensurar a incerteza envolvida no teste em questão. Uma forma de medir essa incerteza é através do cálculo da sensibilidade e especificidade do teste.

A sensibilidade nos expressa a probalilidade de que o teste dê positivo quando de fato o indivíduo apresenta a característica de interesse, e a especificidade nos expressa a probabilidade de que o teste dê negativo quando o indivíduo de fato não apresenta a característica de interesse.

4.3.2 Modelo de regressão logística corrigido

Considere as variáveis Z e Y,tais que Y_i é o resultado do teste de diagnóstico alternativo utilizado.

Logo,

$Y_i = 1$ se o teste do i-ésimo indivíduo der positivo,

$Y_i = 0$ caso contrário,

e Z_i é resultado do teste padrão-ouro, isto é $Z_i = 1$ se o i-ésimo indivíduo realmente apresenta o desfecho e $Z_i = 0$ caso contrário.

Note que na verdade queremos inferir sobre a incógnita Z, no entanto o que temos em mãos é justamente a incógnita Y. Através do teorema da probabilidade total e da

definição de sensibilidade e especificidade podemos encontrar uma relação entre as duas variáveis citadas. Temos que

$$\begin{aligned} P(Y_i = 1) &= P(Y_i = 1|Z_i = 1)P(Z_i = 1) + P(Y_i = 1|Z_i = 0)P(Z_i = 0) \\ &= s\theta + (1 - e)(1 - \theta) \\ &= \theta(s + e - 1) + 1 - e = \kappa(\theta_i, s, e) \end{aligned}$$

onde $s = P(Y_i = 1|Z_i = 1)$ e $e = P(Y_i = 0|Z_i = 0)$ são respectivamente a sensibilidade e a especificidade do teste, e $\theta_i = P(Z_i = 1)$ é a probabilidade do i-ésimo indivíduo realmente apresentar o desfecho.

Parte-se do princípio que os indivíduos são independentes uns dos outros, e assume-se que Y_i segue uma distribuição de Bernoulli com probabilidade $\kappa(\theta_i, s, e)$. Sendo assim, a função de verossimilhança se o tamanho da amostra for n é dada por:

$$L(y; \theta) = \prod_{i=1}^{n} \kappa(\theta_i, s, e)^{y_i}(1 - \kappa(\theta_i, s, e))^{1-y_i}. \tag{4.1}$$

Note que se a sensibilidade e especificidade são iguais a 1, o modelo corrigido de regressão logística e o modelo de regressão logística usual coincidem.

Seguindo com a construção do MLG, é necessário definir qual será a função de ligação utilizada. É natural pensar em

$$g(P(Y = 1)) = \log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = x_i^T\beta$$

. No entanto se a função escolhida for essa estaremos "fugindo"do objetivo de inferir sobre Z, e possivelmente fazendo interpretações erradas. Para não cometer este erro, tem-se que garantir que o objetivo seja satisfeito, ou seja,

$$g(\theta_i) = logit(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) = x_i^T\beta, \tag{4.2}$$

o que implica

$$\theta_i = (logit(\theta_i))^{-1} = \frac{x_i^T\beta}{1 + x_i^T\beta} = g^{-1}(x_i^T\beta). \tag{4.3}$$

Sendo assim substituindo (4.3) em (4.1). Temos que

$$P(Y = 1) = \theta_i(s + e + 1) + 1 - e \tag{4.4}$$

$$= g^{-1}(x_i^T\beta)(s + e + 1) + 1 - e, \tag{4.5}$$

daí tem-se

$$g^{-1}(x_i^T\beta) = \frac{P(Y=1)+e-1}{s+e-1}, \tag{4.6}$$

o que implica

$$x_i^T\beta = g\left(\frac{P(Y=1)+e-1}{s+e-1}\right) = logit\left(\frac{P(Y=1)+e-1}{s+e-1}\right) \tag{4.7}$$

de (4.7), temos a função de ligação que possibilitará fazer análises sobre Z através das informações contidas em Y, s e e.

Lembrando que, $s = P(Y_i = 1|Z_i = 1)$, $e = P(Y_i = 0|Z_i = 0)$,$\theta_i = P(Z_i = 1)$ e $logit(x) = \log(x/1-x)$.

De (4.7), tem-se as seguintes restrições: $P(Y = 1) + e - 1 > 0$ e $s + e - 1 > 0$.

4.3.3 Estimação

As estimativas de β serão obtidas através da equação a seguir. É bom lembrar que a estimativa será feita da mesma forma que em qualquer MLG, com o diferencial de que a função de ligação utilizada será a função logit aplicada a $\kappa(P(Y = 1), s, e)$. Note que, em relação a Y a função de ligação utilizada não é a função logit, mas sim, a que permite que as interpretações do modelo de regressão logística usual sejam feitas em relação a Z.

$$b^{(m)} = (X^TW^{(m-1)}X)^{-1}X^TW^{(m-1)}z^{(m-1)}, \tag{4.8}$$

onde,

$W_{ii} = \frac{(s+e-1)^2}{\kappa(\hat{\theta}_i,s,e)(1-\kappa(\hat{\theta}_i,s,e))}.\left(\frac{d\theta_i}{d\eta_i}\right)^2$, e $z_i = \hat{\eta}_i + (y_i - \kappa(\hat{\theta}_i, s, e)).\left(\frac{d\eta_i}{d\theta_i}\right)$.

5 Estudo simulado

Visando um melhor entendimento sobre os principais fatores que diferenciam o modelo de regressão logística usual do modelo corrigido de regressão logística foi realizado um estudo simulado.

Este estudo foi dividido em duas partes. Na primeira parte o objetivo era conseguir ao mínimo uma intuição sobre em quais casos, de fato, é melhor utilizarmos o modelo corrigido de regressão logística. Entender para qual combinação entre a sensibilidade, especificidade e prevalência o modelo corrigido de regressão logística é melhor do que o modelo usual de regressão logística, e em quais casos não se faz necessário utilizar o modelo corrigido. Sendo assim, esta primeira parte consistiu em gerar p_y (probabilidade do teste do dar positivo) fixando valores para a sensibilidade, especificidade e θ_i na equação (4.1), sendo logit(θ_i)= β_0. Em seguida foi gerada uma amostra de tamanho 100 de uma bernoulli com probablidade p_y.Por fim o parâmetro β_0 e um intervalo de confiança de 95% para o mesmo foram estimados através dos modelos usual e corrigido de regressão logística. As figuras 1, 2 e 3 resumem alguns dos resultados encontrados neste estudo simulado.

Note que quando a sensibilidade e a especificidade do teste são 100% os dois modelos fazem as mesmas estimativas (Figura 1), resultado já esperado, pois uma vez que o teste utilizado é 100% confiável o modelo corrigido de regressão logística não acrescenta nenhuma informação ao modelo usual de regressão logística. Os resultados apresentados nos trazem a intuição de que baixas sensibilidades nos geram estimativas de prevalência menores do que a verdadeira prevalência (com efeito, baixa sensibilidade nos remete a dificuldade de detectar os verdadeiros positivos), de forma análoga baixas especificidades geram estimativas de prevalência maiores do que a verdadeira prevalência.

Nas figuras a seguir é possível verificar que o erro de estimação cometido pelo modelo tradicional de regressão logística pode ser agravado ou atenuado pela sensibilidade ou especificidade dependendo do verdadeiro valor do parâmetro.

Quando $\beta = -1$ (Figura 2), o verdadeiro valor da prevalência é aproximadamente 0,27,

Figura 1: Estimativas pontuais e intervalos de confiança de 95% para β_0 para diferentes valores da sensibilidade e especificidade.

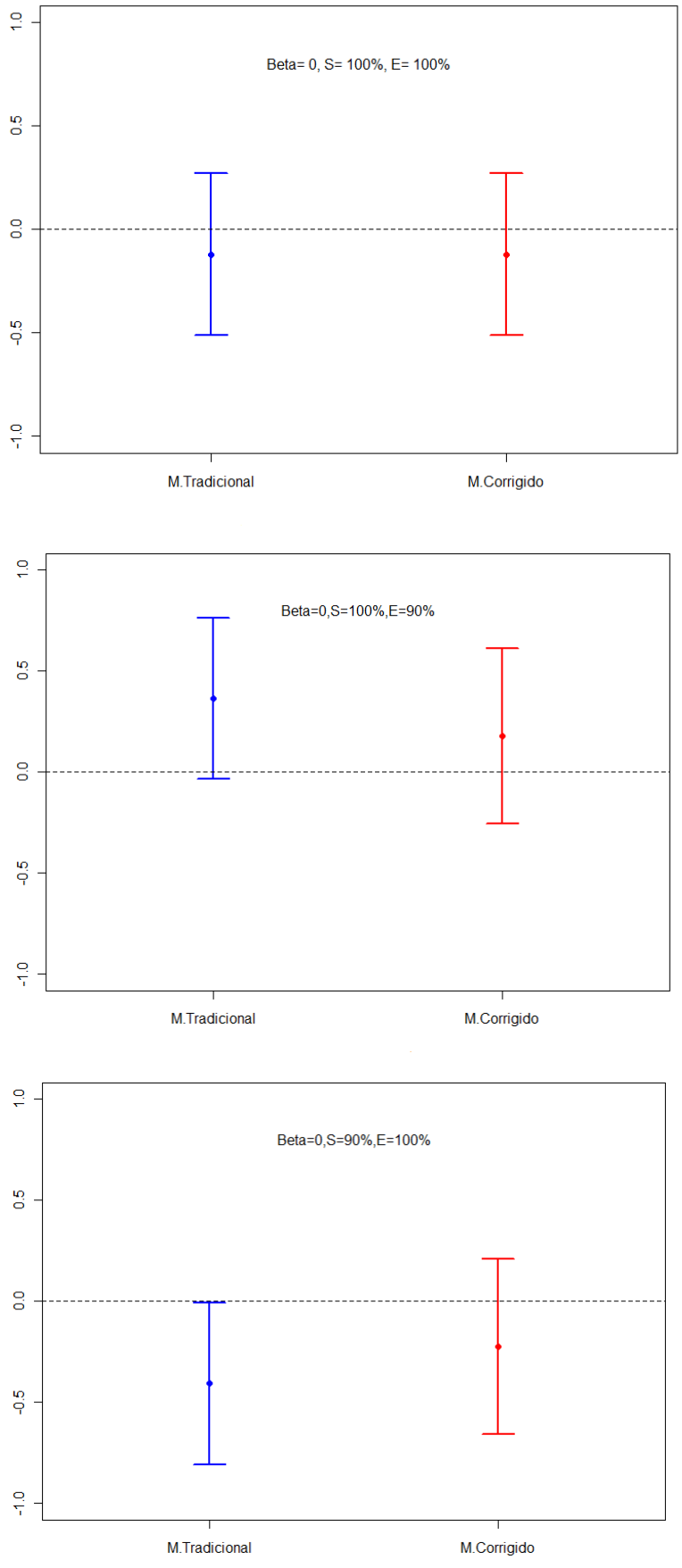


Figura 2: Comparação do peso da sensibilidade e especificidade ($\beta_0 = -1$).

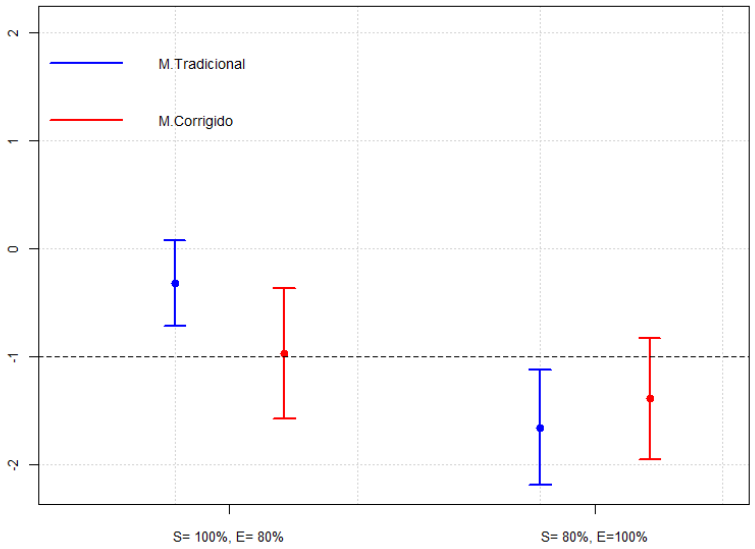
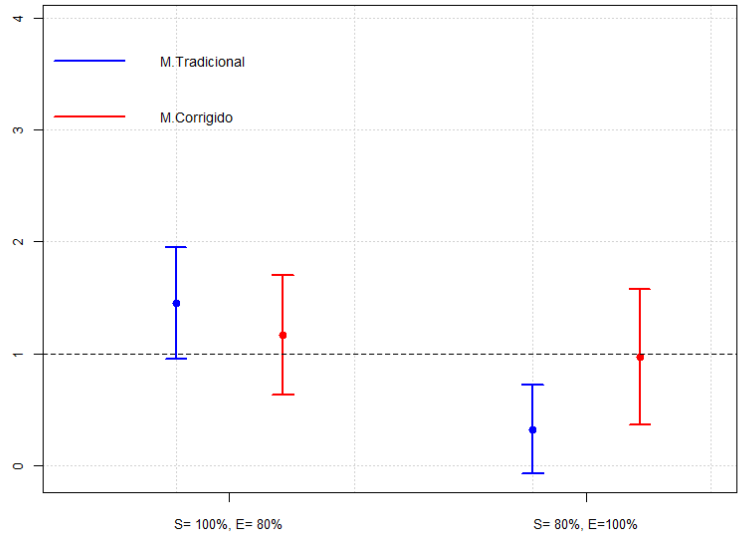


Figura 3: Comparação do peso da sensibilidade e especificidade ($\beta_0 = 1$).



neste caso se faz importante que o valor da especificidade seja alto, pois os indivíduos que não possuem a característica em questão são maioria. No entanto quando $\beta = 1$ (Figura 3), o verdadeiro valor da prevalência é aproximadamente 0,73, neste caso se faz importante que o valor da sensibilidade seja alto, pois os indivíduos que possuem a característica em questão são maioria.

Além do estudo simulado foi realizado um estudo Monte Carlo e nele foi possível verificar que o modelo corrigido manteve-se consistente gerando estimativas pontuais próximas aos verdadeiros valores do parâmetro, e intervalos de confiança com boa cobertura. O modelo tradicional no entanto só apresentou esses resultados para valores de sensibilidade e especificidade muito próximos de 100%.

No estudo Monte Carlo, tendo em mãos a amostra gerada com os resultados dos testes, foi feita a estimação pontual e intervalar do parâmetro através dos dois modelos. Este procedimento foi repetido 1000 vezes. Com o resultado dessas 1000 iterações foi calculado o EQM (erro quadrático médio), a cobertura do intervalo de confiança (percentual de vezes que o verdadeiro valor do parâmetro estava contido no intervalo de confiança) e o vício relativo utilizando as estimativas de cada modelo. As tabelas abaixo mostram o resultado obtido considerando que o modelo possui apenas um parâmetro.

É possível verificar que o modelo corrigido manteve-se consistente gerando estimativas pontuais próximas aos verdadeiros valores do parâmetro, e intervalos de confiança com boa cobertura. O modelo tradicional no entanto só apresentou esses resultados para valores de sensibilidade e especificidade muito próximos de 100%.

Tabela 1: Erro quadrático médio, cobertura do IC95% e vício relativo com base nas estimativas encontradas no estudo de Monte Carlo para diferentes valores de sensibilidade e especificidade.

Sensibilidade= 90% , Especificidade=100%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6615	0,3217	0,1559	0,6300
Corrigido	0,7349	-0,0410	0,0921	0,9550

Sensibilidade=90% , Especificidade=95%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6724	0,2725	0,1235	0,7000
Corrigido	0,7322	-0,0265	0,0897	0,9450

Sensibilidade= 90% , Especificidade=90%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6899	0,1913	0,0839	0,8100
Corrigido	0,7374	-0,0553	0,1003	0,9800

Sensibilidade= 90% , Especificidade=85%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6998	0,1443	0,0690	0,8900
Corrigido	0,7330	-0,0335	0,1034	0,9650

Sensibilidade= 90% , Especificidade=75%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,7254	0,0161	0,0566	0,9350
Corrigido	0,7314	-0,0341	0,1442	0,9600

Sensibilidade= 90% , Especificidade=50%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,7954	-0,3756	0,2020	0,6600
Corrigido	0,7385	-0,1512	0,7742	0,9850

Sensibilidade= 85% , Especificidade=100%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6196	0,5073	0,2984	0,3200
Corrigido	0,7289	-0,0081	0,0830	0,9400

Sensibilidade= 85% , Especificidade=95%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6378	0,4278	0,2294	0,4400
Corrigido	0,7347	-0,0442	0,1107	0,9650

Sensibilidade= 85% , Especificidade=90%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6434	0,4043	0,2014	0,4700
Corrigido	0,7245	0,0115	0,0956	0,9750

Sensibilidade= 85% , Especificidade=85%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6627	0,3173	0,1457	0,6500
Corrigido	0,7324	-0,0371	0,1314	0,9700

Sensibilidade= 85% , Especificidade=75%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6931	0,1760	0,0807	0,8500
Corrigido	0,7384	-0,0832	0,2022	0,9750

Sensibilidade= 85% , Especificidade=50%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,7509	-0,1174	0,0709	0,9300
Corrigido	0,7167	-0,0797	0,9515	0,9450

Sensibilidade= 75% , Especificidade=100%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,5493	0,8006	0,6762	0,0050
Corrigido	0,7324	-0,0305	0,1040	0,9650

Sensibilidade=75% , Especificidade=95%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,5632	0,7438	0,5876	0,0250
Corrigido	0,7331	-0,0384	0,1226	0,9800

Sensibilidade= 75% , Especificidade=90%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,5725	0,7049	0,5392	0,0800
Corrigido	0,7269	-0,0171	0,1699	0,9600

Sensibilidade= 75% , Especificidade=85%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,5886	0,6379	0,4546	0,1450
Corrigido	0,7309	-0,0536	0,2455	0,9600

Sensibilidade= 75% , Especificidade=75%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6151	0,5262	0,3197	0,2500
Corrigido	0,7302	-0,0744	0,3648	0,9650

Sensibilidade= 50% , Especificidade=100%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,3620	1,5747	2,5355	0,0000
Corrigido	0,7240	-0,0422	0,3680	0,9250

Sensibilidade= 50% , Especificidade=95%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,3831	1,4818	2,2418	0,0000
Corrigido	0,7402	-0,2222	1,2472	0,9700

Sensibilidade= 50% , Especificidade=90%				
Modelo	**Prev.est**	**Vicio**	**EQM**	**Cobertura**
Simples	0,6378	0,4278	0,2294	0,4400
Corrigido	0,7347	-0,0442	0,1107	0,9650

6 Aplicação

6.1 Asma

Quando o grupo de interesse são crianças asmáticas, um instrumento utilizado para a identificação deste grupo é o questionário padronizado relacionado a fase I do "International Study of Asthma and Allergies in Childhood" - ISAAC *(Solé et al. 1998)*. A metodologia do ISAAC consiste em 8 questões relacionadas a saúde respiratória de crianças e adolescentes, padronizado para os seguintes grupos de idade: i) crianças de 6 e 7 anos; ii) adolescentes de 13 e 14 anos. As respostas de cada uma das 8 questões são ponderadas e os pontos somados. Adolescentes marcando 6 ou mais pontos são considerados asmáticos, e crianças marcando 5 ou mais pontos são consideradas asmáticas. No Brasil o instrumento foi validado, apresentando a sensibilidade de 92% e 89% para os grupos de 6 e 7, e 13 e 14 anos, respectivamente; e a especificidade de 100% para os dois grupos de estudo *(Solé et al. 1998)*.

Os dados utilizados para essa aplicação correspondem a uma amostra de 574 crianças e adolescentes de 6 a 15 anos de idade residentes no município de Rio Branco, no estado do Acre, e no município de Tangará da Serra, no estado de Mato Grosso, localizados na região da Amazônia Brasileira. A população de estudo foi selecionada aleatoriamente de duas escolas, uma em cada município, respeitando a proporção de alunos por sala de aula. Um questionário estruturado foi respondido pelos responsáveis das crianças e adolescentes. O questionário incluiu o ISAAC (existência ou não de chiado alguma vez na vida ou nos últimos 12 meses, quantidades de crises de chiados nos últimos 12 meses, quantidade de vezes que a criança não dormiu bem por causa da existência de chiado nos últimos 12 meses, e outras) e variáveis envolvidas com tabagismo no domicílio, hábitos alimentares, alergia a produtos de limpeza, proximidade com áreas de queimadas, entre outras.

Considerando que o ISAAC foi padronizado somente para dois grupos específicos de idade (6 e 7, 13 e 14), o banco de dados foi reduzido a esses grupos. Após a redução

do banco, foram calculadas as prevalências de asma, segundo o método ISAAC, para os dois grupos, e modelos de regressão logística simples foram ajustados considerando como variável resposta a presença de asma. Todos os resultados foram estimados com a abordagem usual e corrigida.

6.1.1 Resultados

Na amostra, a prevalência de testes positivos da asma segundo o ISAAC é de 27,4% para o grupo de 6 e 7 anos e 16,7% para o grupo de 13 e 14 anos. Como no ISAAC, a especificidade é de 100%, nos dois grupos de estudo *(Solé et al. 1998)*, a abordagem corrigida $\widehat{P(Z_i = 1)}$ se resume a $\frac{\widehat{P(Y=1)}}{s}$. Logo, a prevalência corrigida é de 29,8% para o grupo de 6 e 7 anos e 18,7% para o grupo de 13 e 14 anos.

As tabelas 2 e 3 apresentam as razões de chance brutas estimadas a partir do modelo de regressão logística usual e corrigido pela sensibilidade e especificidade do ISAAC nos grupos de 6 e 7 anos e 13 e 14 anos, respectivamente. Observa-se em algumas variáveis, como por exemplo 'tosse pela manhã' e 'possuir aparelho nebulizador' aumento nas estimativas pontuais das razões de chance quando utilizamos o modelo corrigido de regressão logística, assim como um aumento na amplitude dos intervalos de confiança.O aumento dos intervalos de confiança quando utiliza-se o modelo corrigido de regressão logística era esperado, uma vez que ao incluir as informações da sensibilidade e especifcidade aumentam-se as incertezas.

Além dos modelos simples foram ajustados também modelos múltiplos para as duas abordagens estudadas. A partir do modelo simples da variável mais siginificativa foi utilizado o método stepwise para seleção das variáveis do modelo, tendo como critério de seleção o teste da razão de verossimilhanças. Primeiramente foi ajustado um modelo com base no método usual, para as 2 faixas-etárias em questão. Em seguida o modelo encontrado foi reproduzido para o modelo corrigido. Como era esperado os resultados encontrados não foram os mesmos (vide tabelas 4 e 6). O modelo de regressão logística corrigido apresentou maior desvio padrão, e em alguns casos os parâmetros estimados foram mais significativos num modelo do que em outro. Por fim realizou-se o mesmo passo a passo utilizado na seleção de variáveis do modelo de regressão logística usual para o modelo de regressão logística corrigido, com o objetivo de verificar se ao utilizar o mesmo método e critério de seleção nas duas abordagens o modelo ajustado seria o mesmo. No banco de dados referente as crianças (6 e 7 anos) encontrou-se modelos diferentes nas duas abordagens, mesmo utilizando-se o mesmo método e critério de seleção de variáveis

Tabela 2: Estimativas das Razões de Chance bruta (OR) e respectivos intervalos de confiança de 95%, segundo a abordagem clássica e a correção, para o diagnóstico de asma (ISAAC) no grupo de crianças de 6 e 7 anos.

Covariáveis	**Abordagem tradicional**			**Abordagem corrigida**		
	OR	**IC95%**	**Valor p**	**OR**	**IC95%**	**Valor p**
Tosse pela manhã, depois de se levantar?						
Sim	3,54	0,95-13,20	0,059	3,79	0,92-15,62	0,065
Não	1.00			1.00		
Faz algum tipo de tratamento para a saúde?						
Sim	4,50	0,68-29,97	0,120	5,02	0,58-43,43	0,1423
Não	1.00			1.00		
Tem aparelho inalador/nebulizador?						
Sim	7,18	1,75-29,38	0.006	8,28	1,65-41,52	0,010
Não	1.00			1.00		
Tem alguma pessoa que fuma ou já fumou na casa?						
Sim, fuma	1.00			1.00		
Sim, parou	1.70	0,45-6,71	0,427	1,73	0,45-6,71	0,427
Não	1,13	0,25-5,12	0,871	1,14	0,24-5,36	0,871
A senhora fumou durante a gravidez da criança?						
Sim	1,09	0,26-4,62	0,911	1,09	0,24-4,87	0,911
Não	1.00			1.00		
Próximo a sua casa ocorre queimadas com fumaça no ambiente?						
Sim	3,40	0,96-12,03	0,058	3,53	0,96-12,92	0,057
Não	1.00			1.00		

(6.1.1). O mesmo não aconteceu para o banco de dados referente aos adolecentes (13 e 14 anos).

Tabela 3: Estimativas das Razões de Chance bruta (OR) e respectivos intervalos de confiança de 95%, segundo a abordagem clássica e a correção, para o diagnóstico de asma (ISAAC) no grupo de crianças de 13 e 14 anos.

Covariáveis	**Abordagem tradicional**			**Abordagem corrigida**		
	OR	**IC95%**	**Valor p**	**OR**	**IC95%**	**Valor p**
Tosse pela manhã, depois de se levantar?						
Sim	3,15	0,89-11,15	0,079	3,24	0,88-12,04	0,078
Não	1.00			1.00		
Faz algum tipo de tratamento para a saúde?						
Sim	3,25	0,92-11,49	0,067	3,35	0,90-12,4	0,070
Não	1.00			1.00		
Tem aparelho inalador/nebulizador?						
Sim	7,18	1,75-29,38	0.006	8,28	1,65-41,52	0,010
Não	1.00			1.00		
Tem alguma pessoa que fuma ou já fumou na casa?						
Sim, fuma	1.00			1.00		
Sim, parou	1,28	0,29-5,63	0,744	1,28	0,29-5,75	0,744
Não	3,29	0,84-12,86	0,086	3,38	0,84-13,59	0,086
A senhora fumou durante a gravidez da criança?						
Sim	0,72	0,14-3,61	0,682	0,71	0,14-3,69	0,682
Não	1.00			1.00		
Próximo a sua casa ocorre queimadas com fumaça no ambiente?						
Sim	1,23	0,39-3,89	0,727	1,23	0,38-3,98	0,727
Não	1.00			1.00		

Tabela 4: Modelo múltiplo - Comparação dos resultados obtidos entre abordagem usual e corrigida no grupo de crianças de 6 e 7 anos - Seleção de variáveis com base na abordagem usual.

	Modelo Tradicional			**Modelo Corrigido**		
	$\boldsymbol{\beta}$	**DP**	**Valor p**	$\boldsymbol{\beta}$	**DP**	**Valor p**
Tem aparelho inalador\nebulizador?	1,703	0,845	0,044	1,708	0,955	0,074
Faz algum tratamento para a saude?	1,214	1,166	0,298	1,207	1,315	0,359
Mora perto de queimadas?	1,375	0,753	0,068	1,446	0,796	0,069
Tosse pela manhã?	1,541	0,824	0,061	1,651	0,922	0,074

Tabela 5: Modelo múltiplo - Comparação dos resultados obtidos entre abordagem usual e corrigida no grupo de crianças de 13 e 14 anos - Seleção de variáveis com base na abordagem usual.

Covariáveis	**Modelo Tradicional**			**Modelo Corrigido**		
	$\boldsymbol{\beta}$	**DP**	**Valor p**	$\boldsymbol{\beta}$	**DP**	**Valor p**
Tem aparelho inalador\nebulizador?	1,560	0,698	0,026	1,698	0,746	0,023
Tosse pela manhã?	1,451	0,691	0,036	1,581	0,734	0,031

Tabela 6: Modelo múltiplo - Seleção de variáveis com base na abordagem corrigida. (grupo de crianças de 6 e 7 anos)

Covariáveis	$\boldsymbol{\beta}$	**Desvio Padrão**	**Valor p**
Tem aparelho inalador\nebulizador?	2,058	0,895	0,021
Mora perto de queimadas?	1,402	0,763	0,066
Tosse pela manhã?	1,426	0,837	0,089

7 Conclusão

O modelo de regressão logística corrigido pode ser muito importante em estudos que fazem o uso de testes diagnósticos. Através desse modelo é possível certificar-se de que as conclusões realizadas estão sendo feitas baseadas na variável que de fato interessa. No entanto, os resultados obtidos através dos modelos usual e corrigido nem sempre são tão diferentes. Em estudos que fazem o uso de testes diagnósticos como o ISAAC, ou seja, com altos valores de sensibilidade e especificidade, os modelos chegam a coincidir (no que tange a escolha das variáveis explicativas). Em contrapartida, conforme foi visto no estudo simulado o erro cometido quando os valores de sensibilidade e especificidade não são extremamente altos podem ser muito significativos. Como a estimação e a inferência do modelo de regressão logística corrigido, pode ser facilmente realizada através dos métodos de MLG, a utilização deste modelo, sempre que possível, pode ser uma boa alternativa.

8 Referências Bibliográficas

Solé, D. et al. ***International Study of Asthma and Allergies in Childhood (ISAAC) written questionnaire: validation of the asthma component among Brazilian children****,Journal of Investigational Allergology & Clinical Immunology,vol.8,no.6,376-382, 1998*

McCullagh, P. and Nelder, J. ***Generalized Linear Models****,Chapman and Hal,1989*;

Jannuzzi, P.M., ***Indicadores sociais no Brasil: conceitos, fonte de dados e aplicações****,Alínea,2001*;

Van der Hout, A. and Van der Heijden, P.G.M and Gilchrist, R., ***The logistic regression model with response variables subject to randomized response****,Computational Statistics and Data Analysis,no.51,6060 – 6069,2007*;

Dobson, A.J, ***An introduction to generalized linear models****,Chapman and Hall,2002*;

Gujarati, D.N., ***Econometria básica****, Bookman,2011*;

Demetrio, C.G.B., ***Modelos Lineares Generalizados em Experimentação Agronômica****,2002*

Fahrmeir, L. and Kaufman, H., ***Consistency and asymptotic normality of maximum likelihood estimator in generalized linear models****, Annals of Statistics, 13, 342-68, 1985*;

Gauss, M. C. and Clarice, G.B.D., ***Modelos Lineares Generalizados****,2007*;

Mastella, G. , ***Teste do suor: a análise de condutividade pode tomar o lugar do método clássico de Gibson e Cooke****, Jornal de Pediatria, vol.86 no.2,2010*

9 Anexo - Implementação no R

```
# Encontrando o parametro p para gerar os dados:

py<-function(s,e,beta){
  g.beta<- exp(beta)/(1+exp(beta))
  prob.y<- s*(g.beta) + (1-e)*(1-g.beta)
  return(prob.y)
}


#####INTERVALO DE CONFIANÇA


#x é o modelo
#Essa função vai nos retornar o beta estimado, e o seu intervalo de confiança

Int_beta<-function(x){
  aux<-summary(x)
  beta.est<-aux$coefficients[1,1]
  sd.est<-aux$coefficients[1,2]
  int.inf<-(beta.est-(1.96*sd.est))
  int.sup<-(beta.est+(1.96*sd.est))
  ret<-data.frame("Beta_estimado"=beta.est,"LI_IC95"=int.inf,"LS_IC95"=int.sup)
  return(ret)
}


#Estimando a prevalencia

prev<-function(x){
  beta0=x$coefficients[1]
  p=(exp(beta0)/(1+exp(beta0 )))
  return(p)
}


### FUNÇÕES
```

```
#FUNÇÃO INDICADORA \ COBERTURA
#(indica se o beta verdadeiro realmente está contido no intervalo estimado)

#beta é o beta estimado
#int é o intervalo de confiança com base no beta estimado

INDIC= function(beta,int){
  LI = int[1]
  LS = int[2]
  answer = ifelse( LI <= beta, ifelse( LS > beta, 1, 0 ),  0)
  answer = as.numeric(answer)
  return(answer)
}

# VIÉS
# Vai me retornar o viés relativo

BIAS = function(beta, est){
  answer= drop((beta-est)/beta)
  answer= as.numeric(answer)
  return(answer)

}

#EQM
#Retorna o erro quadrático

EQM = function(beta, est){
  answer= drop((beta-est)^2)
  answer= as.numeric(answer)
  return(answer)
}

#MONTE CARLO:
#Esta função retorna o beta medio,prevalencia media,
vicio relativo, eqm, e cobertura em K repetições

MC= function(beta,s,e,K){

  # REGRESSÃO LOGISTICA SIMPLES
  media.s = 0
  vicio.s = 0
  cobertura.s = 0
  eqm.s = 0
  sprev.s=0

  # REGRESSÂO LOGISTICA CORRIGIDA
  media.c = 0
```

```
vicio.c = 0
cobertura.c = 0
eqm.c = 0
sprev.c=0

banco=data.frame(y=rep(0,100))

for(k in 1:K){

  # Gerando os dados:
  p<-py(s,e,beta)
  banco$y<-rbinom(100,1,p)


  #Estimando beta e o intervalo:
  #modelo simples
  prev.s<-mean(banco$y)
  mod<-glm(y~1,family=binomial(link=logit), data=banco)
  Int.s<-Int_beta(mod)
  #modelos corrigido0
  mod_c<-logreg.es(y ~ 1, esp=e, sen=s, data=banco, new.rap=T)
  prev.c<-prev(mod_c)
  Int.c<-Int_beta(mod_c)

  #Calculando a soma das prevalências
  sprev.s=sprev.s +prev.s
  sprev.c=sprev.c +prev.c

  # Calculando a soma do viés:
  vicio.s = vicio.s+ BIAS(beta,Int.s[1])
  vicio.c = vicio.c+ BIAS(beta,Int.c[1])

  #Calculando a soma do EQM
  eqm.s= eqm.s+EQM(beta,Int.s[1])
  eqm.c= eqm.c+EQM(beta,Int.c[1])

  #Calculando a soma da cobertura
  cobertura.s=cobertura.s+INDIC(beta,Int.s[2:3])

  cobertura.c=cobertura.c+INDIC(beta,Int.c[2:3])


  #print(c(k,cobertura.s, cobertura.c))
  #print(Int.s[2:3])
  #print(Int.c[2:3])

  if( k %% 50 == 0)
    print(k)
```

```
  }
  # Vicio
  vicio.s=vicio.s/K
  vicio.c=vicio.c/K

  # EQM
  eqm.s=eqm.s/K
  eqm.c=eqm.c/K

  #Cobertura

  cobertura.s=cobertura.s/K
  cobertura.c=cobertura.c/K

  #Prevalencia
  sprev.s=sprev.s/K
  sprev.c=sprev.c/K

  #Criando um data.frame com os resultados:

  saida=data.frame("Prev.est"=c(sprev.s,sprev.c),"Vicio"=c(vicio.s,vicio.c),"EQM"=c
                   "Cobertura"=c(cobertura.s,cobertura.c),row.names=c("Simples","Co
  return(saida)

}

#SIMPLES:
mod<-glm(y~1,family=binomial(link=logit), data=banco)
#CORRIGIDO:
mod_c<-glm(y ~ 1, family=binomial(logitse(s,e)), data=banco)

####ESTIMANDO A PREVAL?NCIA:

#MODELO SIMPLES:

prev.s<-mean(y)

#MODELO CORRIGIDO;

prev.c<-prev(mod_c)

####INTERVALO DE CONFIAN?A:

Int_beta_s(mod)

Int_beta_c(mod_c)
```

```
####MONTE CARLO

#Simulando para Beta=1

beta=1
K=200

#sensibilidade=100
s=1

#especificidade=100, 95, 90, 85, 75, 50

e=1
s1=MC(beta,s,e,K)

e=.95
s2=MC(beta,s,e,K)

e=.90
s3=MC(beta,s,e,K)

e=.85
s4=MC(beta,s,e,K)

e=.75
s5=MC(beta,s,e,K)

e=.5
s6=MC(beta,s,e,K)


#sensibilidade=95
s=.95

#especificidade=100, 95, 90, 85, 75, 50

e=1
s7=MC(beta,s,e,K)

e=.95
s8=MC(beta,s,e,K)

e=.90
s9=MC(beta,s,e,K)

e=.85
s10=MC(beta,s,e,K)
```

```
e=.75
s11=MC(beta,s,e,K)

e=.5
s12=MC(beta,s,e,K)

#sensibilidade=90
s=.90

#especificidade=100, 95, 90, 85, 75, 50

e=1
s13=MC(beta,s,e,K)

e=.95
s14=MC(beta,s,e,K)

e=.90
s15=MC(beta,s,e,K)

e=.85
s16=MC(beta,s,e,K)

e=.75
s17=MC(beta,s,e,K)

e=.5
s18=MC(beta,s,e,K)

#sensibilidade=85
s=.85

#especificidade=100, 95, 90, 85, 75, 50

e=1
s19=MC(beta,s,e,K)

e=.95
s20=MC(beta,s,e,K)

e=.90
s21=MC(beta,s,e,K)

e=.85
s22=MC(beta,s,e,K)

e=.75
s23=MC(beta,s,e,K)
```

```
e=.5
s24=MC(beta,s,e,K)

#sensibilidade=75
s=.75

#especificidade=100, 95, 90, 85, 75, 50

e=1
s25=MC(beta,s,e,K)

e=.95
s26=MC(beta,s,e,K)

e=.90
s27=MC(beta,s,e,K)

e=.85
s28=MC(beta,s,e,K)

e=.75
s29=MC(beta,s,e,K)

e=.5
s30=MC(beta,s,e,K)

#sensibilidade=50
s=.50

#especificidade=100, 95, 90, 85, 75, 50

e=1
s31=MC(beta,s,e,K)

e=.95
s32=MC(beta,s,e,K)

e=.90
s33=MC(beta,s,e,K)

e=.85
s34=MC(beta,s,e,K)

e=.75
s35=MC(beta,s,e,K)

e=.5
```

```
s36=MC(beta,s,e,K)

#Primeiro grupo

#Beta=1, s=100, e=85

beta=1
s=1
e=.80

#Gerando os dados

p<-py(s,e,beta)
y<-rbinom(100,1,p)
banco=data.frame(y)

#Modelos

#SIMPLES:
mod1<-glm(y~1,family=binomial(link=logit), data=banco)
#CORRIGIDO:
mod1_c<-glm(y ~ 1, family=binomial(logitse(s,e)), data=banco)

#Intervalo de confiança

b_s1.1<-as.numeric(Int_beta_s(mod1))

b_c1.1<-as.numeric(Int_beta_c(mod1_c))

beta1<-c(b_s1.1[2],b_s1.1[1],b_s1.1[3],b_c1.1[2],b_c1.1[1],b_c1.1[3])

#Segundo grupo


beta=1
s=.80
e=1

#Gerando os dados

p<-py(s,e,beta)
y<-rbinom(100,1,p)
banco=data.frame(y)

#Modelos

#SIMPLES:
mod1<-glm(y~1,family=binomial(link=logit), data=banco)
```

```
#CORRIGIDO:
mod1_c<-glm(y ~ 1, family=binomial(logitse(s,e)), data=banco)

#Intervalo de confiança

b_s1.2<-as.numeric(Int_beta(mod1))

b_c1.2<-as.numeric(Int_beta(mod1_c))

beta2<-c(b_s1.2[2],b_s1.2[1],b_s1.2[3],b_c1.2[2],b_c1.2[1],b_c1.2[3])

#Beta 1

be1<-data.frame(beta1,beta2)
be_1<-data.frame(beta3,beta4)

#Beta -1

be1<-t(be1)
be_1<-t(be_1)



#A funçao do grafico

plots <- function(table, main = "", xlab = "", ylab = "", col = c("darkgreen",
                                                                  "red")) {
  n <- dim(table)[1]  # numero de grupos

  plot(NA, xlim = c(1 - 0.3, n + 0.5), ylim = c(min(table), max(table)+2), main = m
       xlab = xlab, ylab = ylab, axes = F)  # plot vazio com dimensão que varia

  grid()
  box()
  # desenha o quadrado ao redor do gráfico e o grid das linhas pontilhadas

  axis(2, cex.axis = 0.8)  # insere eixo y
  axis(1, seq(1.15, n + 0.15, 1), paste("Caso",1:2), cex.axis = 0.8)
  # Nome dos caras que aparecem no eixo x
  #axis(1, seq(1.15, n + 0.15, 1), rep("",n), cex.axis = 0.8)
   # Nome dos caras que aparecem no eixo x
  for (i in 1:n) {

    arrows(x0 = i, x1 = i, y0 = table[i,1], y1 = table[i,3], code = 3,
           angle = 90, length = 0.15, lwd = 2, col = col[1])
            # arrows para grupo 1
    points(i, table[i,2], cex = 1, lwd = 2, pch = 19, col = col[1])
    #points(i, table$or1[i], cex = 2, lwd = 5, col = col[1])
```

```
    arrows(x0 = i + 0.3, x1 = i + 0.3, y0 = table[i,4], y1 = table[i,6],
           code = 3, angle = 90, length = 0.15, lwd = 2, col = col[2])
             # arrows para grupo 2
    points(i + 0.3, table[i,5], cex = 1, lwd = 2, pch = 19, col = col[2])
    #points(i + 0.3, table$or2[i], cex = 2, lwd = 5, col = col[2])
  }


}


library(foreign)

#Abrindo o banco

banco<-read.csv2(file="RBTS.csv")
names(banco)
str(banco)
fix(banco)

summary(banco)

#Transformando as variáveis
banco$q98<-as.factor(ifelse(banco$q98==1,1,ifelse(banco$q98==2,0,NA)))
banco$q53<-as.factor(ifelse(banco$q53==1,1,ifelse(banco$q53==2,0,NA)))
banco$q70<-as.factor(ifelse(banco$q70==1,1,ifelse(banco$q70==2,0,NA)))
banco$q72<-as.factor(ifelse(banco$q72==1,1,ifelse(banco$q72==2,0,NA)))
banco$q111<-as.factor(banco$q111)
banco$q125<-as.factor(ifelse(banco$q125==1,1,ifelse(banco$q125==2,0,NA)))
banco$q143<-as.factor(banco$q143)
banco$q144<-as.factor(banco$q144)
banco$q145<-as.factor(ifelse(banco$q145==1,1,ifelse(banco$q145==2,0,NA)))

banco$idade6e7<-as.factor(ifelse(banco$Idade==6,1,ifelse(banco$Idade==7,1,0)))
banco$idade13e14<-as.factor(ifelse(banco$Idade==13,1,ifelse(banco$Idade==14,1,0)))

#OR

OR<-function(x)
{
  beta <- coef(x)
  odds<-exp(beta)
  summ <- summary(x)
  se.beta <- sqrt(diag(summ$cov.unscaled))
  liminf <- exp(beta-1.96*se.beta)
```

```
  limsup <- exp(beta+1.96*se.beta)
  retval<-data.frame("Odds_ratio"=odds,"LI_IC95"=liminf,"LS_IC95"=limsup)
  return(retval)
}

#Analise de residuo

envelope.logistica<-function(mod,rep=100)
{
  X<-model.matrix(mod)
  w<-mod$weights
  W<-diag(w)
  H<-sqrt(W)%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrt(W)
  h<-diag(H)
  n<-nrow(X)
  rd<-resid(mod,type="deviance")
  phi<-1
  td<-rd*sqrt(phi/(1-h))
  re<-matrix(0,n,rep)

  for(i in 1:rep)
  {
    dif<-runif(n)-fitted(mod)
    dif[dif>=0]<-0
    dif[dif<0]<-1
    nresp<-dif
    fit<-glm(nresp~-1+X,family=binomial(link="logit"),maxit=50)
    w1<-fit$weights
    W1<-diag(w1)
    H1<-sqrt(W1)%*%X%*%solve(t(X)%*%W1%*%X)%*%t(X)%*%sqrt(W1)
    h1<-diag(H1)
    re[,i]<-sort(resid(fit,type="deviance")*sqrt(phi/(1-h1)))
  }
  e1<-numeric(n)
  e2<-numeric(n)

  for(i in 1:n)
  {
    eo<-sort(re[i,])
    e1[i]<-eo[ceiling(0.025*rep)]
    e2[i]<-eo[floor(0.975*rep)]
  }

  xb<-apply(re,1,mean)
  limites<-range(td,e1,e2)
  #grafico
  par(pty="s")
  qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=limites,lty=1)
```

```
  par(new=TRUE)
  qqnorm(e2,axes=F,xlab="",ylab="",type="l",ylim=limites,lty=1)
  par(new=TRUE)
  qqnorm(xb,axes=F,xlab="",ylab="",type="l",ylim=limites,lty=2)
  par(new=TRUE)
  qqnorm(td,ylim=limites)
}



##### MODELO SEM CORREÇÃO

# Banco: 6 e 7 anos

banco6e7<-subset(banco, banco$idade6e7==1)


#Modelos Simples (6e7anos)

mod1<-glm(Asma~q98,family=binomial(link=logit), data=banco6e7)
summary(mod1)

mod2<-glm(Asma~q53,family=binomial(link=logit), data=banco6e7)
summary(mod2)

mod3<-glm(Asma~q70,family=binomial(link=logit), data=banco6e7)
summary(mod3)

mod4<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
summary(mod4)

mod5<-glm(Asma~q111,family=binomial(link=logit), data=banco6e7)
summary(mod5)

mod6<-glm(Asma~q125,family=binomial(link=logit), data=banco6e7)
summary(mod6)

mod7<-glm(Asma~q143,family=binomial(link=logit), data=banco6e7)
summary(mod7)


mod8<-glm(Asma~q144,family=binomial(link=logit), data=banco6e7)
summary(mod8)

mod9<-glm(Asma~q145,family=binomial(link=logit), data=banco6e7)
summary(mod9)

#Modelo Multiplo - RAZÃO DA VEROSSIMILHANÇA (nivel de significancia 10%)
```

```
#H0= a nova v.a não é importante (beta=0)
#H1= a nova v.a é importate

ajuste<-function(x,y,z)
{
vazio<-logLik(x)
cheio<-logLik(y)
l<-(-2*(vazio-cheio))
vc<-qchisq(0.1,df=z,lower.tail=FALSE,log.p=FALSE)
des<-ifelse(l > vc,"Rejeita-se H0","Não Rejeita-se H0")
pval<- 1-pchisq(l,z)
retval2<-data.frame("Estatística_l"=l,"Qui_quadrado"=vc,"Decisão"=des,
"P-valor"=pval)
return(retval2)
}

#Fase 1

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q145,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q98,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q111,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q125,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q144,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q70,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

#Fase 2
```

```
mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)


mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q98,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q111,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q125,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q144,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q70,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

# Fase 3

mod<-glm(Asma~q72+q53+q145,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q98,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q145,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q125,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q145,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q144,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q145,family=binomial(link=logit), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q70,family=binomial(link=logit), data=banco6e7)
ajuste(mod,mod1,1)

#MODELO ESCOLHIDO:

mod<-glm(Asma~q72+q53+q145+q98,family=binomial(link=logit), data=banco6e7)
summary(mod)
```

```
# Análise de resíduo

envelope.logistica(mod)


# Banco: 13 e 14 anos

banco13e14<-subset(banco, banco$idade13e14==1)


#Modelos Simples (13e14anos)

mod1<-glm(Asma~q98,family=binomial(link=logit), data=banco13e14)
summary(mod1)

mod2<-glm(Asma~q53,family=binomial(link=logit), data=banco13e14)
summary(mod2)

mod3<-glm(Asma~q70,family=binomial(link=logit), data=banco13e14)
summary(mod3)

mod4<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
summary(mod4)

mod5<-glm(Asma~q111,family=binomial(link=logit), data=banco13e14)
summary(mod5)

mod6<-glm(Asma~q125,family=binomial(link=logit), data=banco13e14)
summary(mod6)

mod7<-glm(Asma~q143,family=binomial(link=logit), data=banco13e14)
summary(mod7)

mod8<-glm(Asma~q144,family=binomial(link=logit), data=banco13e14)
summary(mod8)

mod9<-glm(Asma~q145,family=binomial(link=logit), data=banco13e14)
summary(mod9)



#Modelo Multiplo - RAZÃO DA VEROSSIMILHANÇA (nivel de significancia 10%)

#Fase 1

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q98,family=binomial(link=logit), data=banco13e14)
```

```
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q111,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q125,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q144,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q145,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q143,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q70,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

#Fase 2

mod<-glm(Asma~q72+q98,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q98+q53,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q111,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q144,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q145,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)
```

```
mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q143,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q70,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)


# Fase 3

mod<-glm(Asma~q72+q98,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q98+q145,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q98,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q98+q111,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q98,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q98+q144,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q98,family=binomial(link=logit), data=banco13e14)
mod1<-glm(Asma~q72+q53+q98+q143,family=binomial(link=logit), data=banco13e14)
ajuste(mod,mod1,1)

#MODELO ESCOLHIDO

mod<-glm(Asma~q72+q98,family=binomial(link=logit), data=banco13e14)

# Análise de resíduo

envelope.logistica(mod)


###MODELO COM CORREÇÃO

#Reprodução do modelo encontrado no modelo normal

#Banco: 6 e 7

mod<-glm(Asma~q72+q53+q145+q98,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod)

#Banco 13 e 14

mod<-glm(Asma ~ q72+q98, family=binomial(logitse(.89,1)), data=banco13e14)
```

```
# Passo a passo

# Banco: 6 e 7 anos

banco6e7<-subset(banco, banco$idade6e7==1)


#Modelos Simples (6e7anos) - CORRIGIDO

mod1<-glm(Asma~q98,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod1)

mod2<-glm(Asma~q53,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod2)

mod3<-glm(Asma~q70,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod3)

mod4<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod4)

mod5<-glm(Asma~q111,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod5)

mod6<-glm(Asma~q125,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod6)

mod7<-glm(Asma~q143,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod7)


mod8<-glm(Asma~q144,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod8)

mod9<-glm(Asma~q145,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod9)

#Modelo Multiplo CORRIGIDO- RAZÃO DA VEROSSIMILHANÇA (nivel de significancia 10%)

#H0= a nova v.a não é importante (beta=0)
#H1= a nova v.a é importate

ajuste<-function(x,y,z)
{
  vazio<-logLik(x)
  cheio<-logLik(y)
```

```
  l<-(-2*(vazio-cheio))
  vc<-qchisq(0.1,df=z,lower.tail=FALSE,log.p=FALSE)
  des<-ifelse(l > vc,"Rejeita-se H0","Não Rejeita-se H0")
  pval<- 1-pchisq(l,z)
  retval2<-data.frame("Estatística_l"=l,"Qui_quadrado"=vc,"Decisão"=des, "P-valor"=
  return(retval2)
}

#Fase 1

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q145,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q98,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q111,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q125,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q144,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q70,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

#Fase 2

mod<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)


mod<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q98,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)
```

```
mod<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q111,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q125,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q144,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q70,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

# Fase 3

mod<-glm(Asma~q72+q53+q145,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q98,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q145,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q125,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q145,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q144,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q145,family=binomial(logitse(.92,1)), data=banco6e7)
mod1<-glm(Asma~q72+q53+q145+q70,family=binomial(logitse(.92,1)), data=banco6e7)
ajuste(mod,mod1,1)

#MODELO ESCOLHIDO:

mod<-glm(Asma~q72+q145+q98,family=binomial(logitse(.92,1)), data=banco6e7)
summary(mod)

# Análise de resíduo

envelope.logistica(mod)


# Banco: 13 e 14 anos

banco13e14<-subset(banco, banco$idade13e14==1)
```

```
#Modelos Simples CORRIGIDO (13e14anos)

mod1<-glm(Asma~q98,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod1)

mod2<-glm(Asma~q53,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod2)

mod3<-glm(Asma~q70,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod3)

mod4<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod4)

mod5<-glm(Asma~q111,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod5)

mod6<-glm(Asma~q125,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod6)

mod7<-glm(Asma~q143,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod7)

mod8<-glm(Asma~q144,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod8)

mod9<-glm(Asma~q145,family=binomial(logitse(.89,1)), data=banco13e14)
summary(mod9)



#Modelo Multiplo CORRIGIDO- RAZÃO DA VEROSSIMILHANÇA (nivel de significancia 10%)

#Fase 1

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q98,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q111,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q125,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)
```

```
mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q144,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q145,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q143,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q70,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

#Fase 2

mod<-glm(Asma~q72+q98,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q98+q53,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q111,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q144,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q145,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q143,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q70,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)
```

```
# Fase 3

mod<-glm(Asma~q72+q98,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q98+q145,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q98,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q98+q111,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q98,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q98+q144,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

mod<-glm(Asma~q72+q53+q98,family=binomial(logitse(.89,1)), data=banco13e14)
mod1<-glm(Asma~q72+q53+q98+q143,family=binomial(logitse(.89,1)), data=banco13e14)
ajuste(mod,mod1,1)

#MODELO ESCOLHIDO

mod<-glm(Asma~q72+q98,family=binomial(logitse(.89,1)), data=banco13e14)

# Análise de resíduo

envelope.logistica(mod)

#### FIM####
```