UNIVERSIDADE FEDERAL DE SERGIPE
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS
CURSO DE BACHARELADO EM ESTATÍSTICA

MATHEUS FRANCISCO NASCIMENTO DE JESUS

ESTUDO COMPARATIVO ENTRE AS FUNÇÕES DE LIGAÇÃO LOGIT E PROBIT

Estimando Parâmetros

São Cristovão/SE

2015

MATHEUS FRANCISCO NASCIMENTO DE JESUS

ESTUDO COMPARATIVO ENTRE AS FUNÇÕES DE LIGAÇÃO LOGIT E PROBIT:

Estimando parâmetros

Monografia de conclusão de curso de graduação apresentado à Universidade Federal de Sergipe, como requisito para obtenção do grau de Bacharel em Estatística.

Orientador: Allan Robert da Silva

São Cristovão/SE

2015

MATHEUS FRANCISCO NASCIMENTO DE JESUS

ESTUDO COMPARATIVO ENTRE AS FUNÇÕES DE LIGAÇÃO LOGIT E PROBIT:

Estimando parâmetros

Monografia de conclusão de curso de graduação apresentado à Universidade Federal de Sergipe, como requisito para obtenção do grau de Bacharel em Estatística.

Aprovada em _____ de _____ de _____.

BANCA EXAMINADORA:

_____

Prof. Allan Robert da Silva

_____

Prof. Daniel Francisco Neyra Castaneda

_____

Prof.ª Amanda da Silva Lira

RESUMO

Esse estudo conta com uma boa introdução sobre a família de distribuições exponenciais que é um tema indispensável quando se fala em Modelos Lineares Generalizados e também um bom conteúdo sobre estes últimos. Foram utilizadas as funções de ligação logit e probit com o intuito de avaliar o desempenho das mesmas na estimação de parâmetros com dados alterados. Fixaram-se três valores diferentes para o parâmetro beta, que é uma probabilidade de sucesso, em seguida, foram realizadas 1000 simulações (estimações) para cada combinação dos parâmetro e das funções de ligação. Foi utilizado o Erro Quadrático Médio e a porcentagem de erros quadráticos menores que 0,02 para serem utilizados para comparação. Chegou-se a conclusão de que a função de ligação probit foi a mais indicada para estimar os parâmetros com valores próximos a zero e que a função de ligação logit foi a mais indicada para a estimação de parâmetros com valores próximos a um. Todo o procedimento foi realizado através do software livre R.

Palavras-chave: Modelos Lineares Generalizados. Software R. Simulação.

ABSTRACT

This study has a good introduction to the family of exponential distributions which is an indispensable subject when it comes to Generalized Linear Models and also good content on the latter. The logit and probit link functions in order to evaluate the performance of the same in the estimation of parameters with changed data were used. Three different fixed values for the parameter beta, which is a probability of success, then simulations were performed in 1000 (estimates) for each combination of parameters and connection functions. We used the mean squared error and the percentage of children squared errors than 0.02 to be used for comparison. Came to the conclusion that the probit link function was indicated to estimate the parameters next to zero and that the logit link function was the most suitable for the estimation of parameters next to one. The whole procedure was performed using the free software R.

Keywords: Generalized Linear Models. Software R. Simulation.

SUMÁRIO

1 INTRODUÇÃO

Família Exponencial

A família exponencial de distribuições foi proposta por Darmois, Koopman e Pitman ao estudarem as propriedades da suficiência estatística. Depois, vários outros aspectos dessa família foram estudados e passaram a ser fundamentais na teoria moderna de Estatística. Fisher foi o responsável por introduzir a definição de família exponencial na estatística, mas essa família de distribuições surgiu na Mecânica Estatística no final do século XIX e foi desenvolvida por Boltzmann, Gibbs e Maxwell. Por meio do trabalho inovador de Nelder e Wedderburn, em 1972, que definiu os Modelos Lineares Generalizados (MLG) foi que a importância da família exponencial teve maior destaque nos modelos de regressão.

Uma das poucas condições necessárias para que seja possível a utilização de Modelos Lineares Generalizados é que a variável dependente pertença à família exponencial de distribuições, por isso é indispensável à abordagem a essa família especifica de distribuições antes de introduzirmos os Modelos Lineares Generalizados (MLG). A importância da família exponencial para os MLG's se deve ao fato de ser possível a utilização de dados discretos ou contínuos, dados que possuem assimetria e dados que são restritos a um intervalo do conjunto dos reais, como é o caso da distribuição binomial que tem intervalo (0,1).

As distribuições biparamétricas que pertencem à família exponencial são as que possuem função densidade (ou de probabilidade) que podem ser representadas da seguinte forma:

$$f(y;\theta;\phi) = exp\left\{\frac{y * \theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Em que,

y é a variável de interesse

θ é o parâmetro de localização

ϕ é o parâmetro de dispersão, mais conhecido como σ^2

$a(\phi), b(\theta)\ e\ c(y,\phi)$ são funções denominadas de funções específicas

A média e a variância para as distribuições biparamétricas podem ser obtidas, respectivamente, através da primeira e segunda derivadas da função $b(\theta)$.

$$E(y) = b'(\theta) \qquad Var(y) = a(\phi) * b''(\theta)$$

Caso o parâmetro ϕ seja conhecido, tornando-se uma distribuição uniparamétrica, a família exponencial também pode ser descrita na forma:

$$f(y;\theta) = h(y) * exp\{y * \theta - b(\theta)\}$$

Em que,

y é a variável de interesse

θ é o parâmetro de localização

$b(\theta)\ e\ h(x)$ são funções denominadas de funções específicas

A média e a variância para as distribuições uniparamétricas podem ser obtidas, respectivamente, através da primeira e segunda derivadas da função $b(\theta)$.

$$E(y) = b'(\theta) \qquad Var(y) = b''(\theta)$$

Existem várias distribuições que pertencem à família exponencial, mas utilizaremos como exemplo as mais conhecidas são elas:

- Normal (Geralmente utilizada para dados contínuos simétricos)

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} * exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} = exp\left\{\frac{y*\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2} * \ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right\}$$

Em que,

$$\theta = \mu \ ;\ b(\theta) = \frac{\mu^2}{2} \ ;\ a(\phi) = \sigma^2 \ ;\ c(y,\phi) = \frac{1}{2} * \ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$$

E a média e a variância de y são:

$$E(y) = b'(\theta) = \mu$$

$$Var(y) = b''(\theta) * a(\phi) = 1 * \sigma^2 = \sigma^2$$

- Binomial (Geralmente utilizada para dados de proporções)

$$f(y;\mu) = \binom{n}{y} * p^y * (1-p)^{n-y} = exp\left\{y * ln\left(\frac{p}{1-p}\right) + n * \ln(1-p) + ln\binom{n}{y}\right\}$$

Em que,

$\theta = \ln\left(\frac{p}{1-p}\right) = \left[p = \frac{\mu}{n}\right] = ln\left(\frac{\mu}{n-\mu}\right)$; $b(\theta) = -n * \ln(1-p) = n * \ln(1 + e^{\theta})$; $a(\phi) = 1$; $c(y,\phi) = ln\binom{n}{y}$

E a média e a variância de y são:

$$E(y) = b'(\theta) = \frac{n*e^{\theta}}{1+e^{\theta}} = \left[e^{\theta} = \left(\frac{p}{1-p}\right)\right] = n * p$$

$$Var(y) = b''(\theta) * a(\phi) = \frac{n * e^{\theta}}{(1+e^{\theta})^2} = \left[e^{\theta} = \left(\frac{p}{1-p}\right)\right] = n * p * (1-p)$$

- Poisson (Geralmente utilizada para dados de contagens)

$$f(y;\lambda) = \frac{\lambda^y * e^{-\lambda}}{y!} = exp\{y * ln(\lambda) - \lambda - ln(y!)\}$$

Em que,

$\theta = \ln(\lambda) \rightarrow \lambda = e^{\theta}$; $b(\theta) = \lambda = e^{\theta}$; $a(\phi) = 1$; $c(y,\phi) = -ln(y!)$

E a média e a variância de y são:

$$E(y) = b'(\theta) = e^{\theta} = \lambda$$

$$Var(y) = b''(\theta) * a(\phi) = e^{\theta} * 1 = \lambda$$

- Gama (Geralmente utilizada para dados contínuos assimétricos)

$$f(y;\mu,\alpha) = \frac{1}{\Gamma(\alpha)} * \left(\frac{\alpha}{\mu}\right) * y^{\alpha-1} * e^{-\alpha\left(\frac{y}{\mu}\right)}$$

$$= exp\left[\frac{y * \frac{-1}{\mu} + \ln\left(\frac{1}{\mu}\right)}{\frac{1}{\alpha}} + \alpha * \ln(\alpha * y) - \ln(y) - \ln[\Gamma(\alpha)]\right]$$

Em que,

$$\theta = -\frac{1}{\mu}\ ; b(\theta) = -\ln\left(\frac{1}{\mu}\right) = -\ln(-\theta)\ ;\ a(\phi) = \frac{1}{\alpha}\ ;$$
$$c(y,\phi) = \alpha * \ln(\alpha * y) - \ln(y) - \ln\left[\Gamma(\alpha)\right]$$

E a média e a variância de y são:

$$E(y) = b'(\theta) = -\frac{1}{\theta}$$
$$Var(y) = b''(\theta) * a(\phi) = \mu^2 * \frac{1}{\alpha} = \frac{\mu^2}{\alpha}$$

- Normal Inversa (Geralmente utilizada para dados contínuos assimétricos)

$$f(y;\mu,\sigma^2) = -\frac{1}{\sqrt{2\pi\sigma^2 * y^3}} * exp\left\{\frac{-(y-\mu)^2}{2\sigma^2 * \mu^2 * y}\right\}$$
$$= exp\left\{y * \frac{\frac{-1}{2\mu^2} + \frac{1}{\mu}}{\sigma^2} - \frac{1}{2} * \left[\ln(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 * y}\right]\right\}$$

Em que,

$$\theta = -\frac{1}{2\mu^2}\ ;\ b(\theta) = \frac{1}{\mu} = -(-2 * \theta)^{1/2}\ ;\ a(\phi) = \sigma^2\ ;\ c(y,\phi)$$
$$= -\frac{1}{2} * \left[\ln(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 * y}\right]$$

E a média e a variância de y são:

$$E(y) = b'(\theta) = (-2\theta)^{-1/2} \qquad Var(y) = b''(\theta) * a(\phi) = \mu^3 * \sigma^2$$

Essas são as distribuições mais comumente citadas quando nos referimos a família exponencial de distribuições.

Modelos Lineares Generalizados - MLG

Em um modelo linear clássico a variável resposta é formada basicamente por dois componentes, um sistemático e outro aleatório.

$$Y = \mu + \varepsilon$$

Em que,

Y é o vetor coluna das variáveis dependentes de dimensão nx1,

$\mu = X\beta$, é o componente sistemático,

X é a matriz, de dimensões nxp

β é o vetor coluna dos parâmetros com p elementos,

ε é o vetor coluna de componentes aleatórios com n elementos e $\varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$.

Segundo Demétrio (2001), Nem sempre os requisitos necessário para que essa estrutura seja válida são satisfeitos, tornando assim a utilização do modelo linear clássico inviável. E caso seja utilizada a técnica de regressão clássica sem verificar os devidos pressupostos, isso levará a obtenção de estimativas errôneas a respeito dos parâmetros.

Fora isso, Demétrio (2001) afirma: “Não há razão para restringir à estrutura simples dada por $\mu = E(Y) = X\beta$ para o componente sistemático e nem para se restringir à distribuição normal para o componente aleatório e à suposição de homogeneidade de variâncias.”.

Diante das limitações do modelo clássico de regressão, Nelder e Wedderburn criaram, em 1972, uma técnica estatística que contornava as limitações impostas para o modelo anterior e nomearam de Modelos Lineares Generalizados (MLG), em referencia ao fato de terem utilizado varias técnicas estatísticas em conjunto para fundamentar a teoria dessa nova técnica (por isso o termo “Generalizado”).

Já Cordeiro (1986) afirma que o termo “Generalizado” faz referencia ao fato da variável resposta possuir um leque maior de distribuições e não só a normal como nos modelos lineares clássicos.

Nelder & Wedderburn (1972) demonstraram que a maioria dos problemas estatísticos, que surgem nas áreas de agricultura, demografia, economia, geologia, medicina, ciência política, psicologia, sociologia, zootecnia etc, podem ser

formulados, de uma maneira única, como modelos de regressão. Esses modelos são compostos de uma variável dependente univariada, variáveis explicativas e uma amostra aleatória de n observações e obedecendo as seguintes condições:

1) A variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família exponencial na forma canônica (distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens);
2) As variáveis explicativas entram na forma de um modelo linear (componente sistemático);
3) A ligação entre os componentes aleatório e sistemático é feita através de uma função, denominada função de ligação (por exemplo, logit, probit, complemento log-log).

Estrutura do MLG

Um MLG é formado por três componentes fundamentais, são eles, um Componente aleatório associado a variável resposta, um componente sistemático linear nos parâmetros, denominado preditor linear ou estrutura linear do modelo e uma função de ligação, a qual combina o componente aleatório e o componente sistemático.

- Componente Aleatório

Representado por um conjunto de variáveis aleatórias independentes $Y_1, \dots, Y_n$ obtidas de uma mesma distribuição que faz parte da família de distribuições exponenciais com médias $\mu_1, \dots, \mu_n$, ou seja,

$$E(Y_i) = \mu_i, i = 1, \dots, n.$$

- Componente Sistemático

As variáveis explanatórias entram na forma de uma soma linear de seus efeitos

$$\eta_i = \sum_{r=1}^{p} x_{ir} * \beta_r = x_i^T * \beta \; ou \; \eta = X * \beta$$

Em que,

X é a matriz do modelo, β é o vetor dos parâmetros desconhecidos e η é o preditor linear.

- Função de Ligação

É uma função que relaciona o componente aleatório ao componente sistemático, ou seja, vincula a média ao preditor linear

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

Sendo g(.) uma função monótona e diferenciável. Vejamos alguns exemplos:

- ❖ Identidade: $\eta = \mu$
- ❖ Potência: $\eta = \mu^{\lambda}$, onde λ é um número real qualquer.
- ❖ Logit: $\eta = \log\left(\frac{\mu}{1-\mu}\right)$
- ❖ Probit: $\eta = \Phi^{-1}(\mu)$
- ❖ Complemento log-log: $\eta = \log[-\log(1-\mu)]$
- ❖ Logaritmo: $\eta = \log(\mu)$

Quando as funções de ligação fornecem estatísticas suficientes são denominadas funções de ligação canônicas. Os modelos, Normal, Poisson, Binomial, Gamma e Normal Inversa têm por funções canônicas, respectivamente,

$$\eta = \mu\ , \qquad \eta = \log(\mu)\ , \qquad \eta = \log\left(\frac{\mu}{1-\mu}\right), \qquad \eta = \mu^{-1}, \qquad \eta = \mu^{-2}$$

<u>Estimação do modelo</u>

- A função de máxima verossimilhança

A solução das equações de máxima verossimilhança é equivalente a uma iteração do método de mínimos quadrados ponderados com uma função de peso

$$w = \frac{\left(\frac{d\mu}{dY}\right)^2}{V}$$

e uma variável dependente modificada

$$y = Y + (z-\mu)/(d\mu/dY),$$

Onde μ, Y e V são baseadas nas estimativas atuais.

Na prática, podemos obter um bom procedimento de partida para a interação da seguinte forma: tomar como uma primeira aproximação $\mu = z$ e calcular Y a partir dele; em seguida calcular w como anteriormente e definir y = Y. Em seguida obter a primeira aproximação para os $\beta's$ por regressão. Pode ser necessário alterar ligeiramente o método para lidar com valores extremos de z.

Por exemplo, com a distribuição binomial provavelmente será suficiente para substituir as instâncias de z = 0 ou z = n por z=1/2 e z=n-(1/2), onde, por exemplo, com as transformações probit e logit $\mu = 0$ ou $\mu = n$ conduziria a valores infinitos para Y.

- Estatística suficiente

Um caso importante ocorre quando θ, o parâmetro da distribuição do componente aleatório, e Y o valor previsto do modelo linear, coincidirem. Então,

$$L = z * Y - g(Y) + h(z) \text{ e } \frac{\partial L}{\partial \beta_i} = \alpha(\phi) * (z - \mu) * x_i$$

As equações de máxima verossimilhança são desta forma $\sum_k (z - \hat{\mu}) * x_{ik} = 0$, o somatório esta sobre as observações. Assim temos,

$$\sum_k z_k * x_{ik} = \sum_k \hat{\mu}_k * x_{ik}$$

Quando θ é também a média da distribuição, isto é, $\mu = \theta = Y$, temos um modelo linear clássico com erro normal, para $g'(\theta) = \theta$ temos,

$$g(\theta) = \frac{1}{2} * \theta^2 + const$$

que determina exclusivamente a distribuição como normal com variância $\frac{1}{\alpha(\phi)}$. A subclasse de modelos para os quais existam estatísticas suficientes, observados anteriormente por Cox e Dempster, estendeu-o para incluir diversas variáveis dependentes.

- Analise do desvio

Um modelo linear é dito ser ordenado se a estimação dos $\beta's$ é feita na mesma sequência da sua declaração no modelo. O fato de ser ordenado (ou parcialmente ordenado) pode ser deduzido da estrutura do modelo; por exemplo, não faz sentido ajustar um termo de interação $(ab)_{ij}$ antes de ajustar os efeitos correspondentes a_i e b_j. Podem também ser implícitos os objetivos da estimação, ou

seja, se uma tendência deve ser removida primeiro, antes da estimação dos demais efeitos.

Mais comumente, no entanto, a ordenação é em certa medida arbitrária, e isso dá origem a problemas difíceis de inferência que não vamos tentar resolver aqui. Para facilitar a exposição das ideias básicas devemos assumir que o modelo considerado é ordenado, e será ajustado sequencialmente termo a termo. Os objetivos do ajustamento serão avaliar quantos termos são necessários para uma descrição adequada dos dados, calcular as estimativas dos parâmetros associados e sua matriz de informação.

Dois modelos extremos são concebíveis para qualquer conjunto de dados, o modelo mínimo que contém o menor conjunto de termos que o problema permite, e o modelo completo no qual todos os Y's são diferentes e combinam os dados completamente, de modo que $\hat{\mu} = z$.

Um caso extremo do modelo mínimo é o modelo nulo, que é equivalente ao ajustamento da média geral apenas e eficazmente consigna toda a variação dos dados para o componente aleatório do modelo, enquanto o modelo completo ajusta exatamente e assim consigna toda a variação dos dados para a parte sistemática. O processo de ajustamento do modelo com um modelo ordenado consiste, assim, de se proceder a uma distância adequada do modelo mínimo relativamente ao modelo completo. Em cada etapa, nós aumentamos a bondade do ajuste para o conjunto atual de dados contra o aumento da complexidade do modelo.

O ajustamento dos parâmetros em cada etapa é feita pela maximização da verossimilhança para o modelo atual e a correspondência entre o modelo e os dados será medido quantitativamente pela formula $-2 * L_{max}$ que propomos chamar de desvio. Para as quatro distribuições especiais têm-se os seguintes desvios:

Normal: $\frac{\sum(z-\hat{\mu})^2}{\sigma^2}$,

Poisson: $2 * \left\{\sum z * \ln\left(\frac{z}{\hat{\mu}}\right) - \sum(z-\hat{\mu})\right\}$,

Binomial: $2 * \left[\sum z * \ln\left(\frac{z}{\hat{\mu}}\right) + \sum(n-z) * \ln\left(\frac{(n-z)}{(n-\hat{\mu})}\right)\right]$,

Gamma: $2 * p * \left[-\sum \ln\left(\frac{z}{\hat{\mu}}\right) + \sum \frac{(z-\hat{\mu})}{\hat{\mu}}\right]$

Note-se que o desvio é medido a partir do modelo completo, de modo que os termos envolvendo constantes, os dados, ou o fator de escala são omitidos. O

segundo termo nas expressões das distribuições Poisson e Gamma é geralmente idêntico a zero.

Associado a cada modelo a um valor r denominado de graus de liberdade que é dado pelo posto da matriz X, ou de forma equivalente, o número de parâmetros linearmente independentes de ser estimado. Para uma amostra de n observações independentes, o desvio para o modelo tem graus de liberdade dos resíduos (n-r).

Os graus de liberdade, multiplicados, se necessário, por um fator de escala, formam uma escala de um conjunto de modelos sequenciais com desvios que possam ser comparados; quando (graus de liberdade residuais X fator de escala) é aproximadamente igual para o desvio do modelo atual, então é improvável um maior ajustamento de componentes sistemáticos.

O fator de escala pode ser conhecido (por exemplo, unidade para a distribuição de Poisson) ou desconhecido (por exemplo, a distribuição normal com variância desconhecida). Se desconhecido pode ser estimável diretamente, por exemplo, por observações repetidas, ou indiretamente do desvio após um modelo adequado ter sido ajustado (estimado).

A adequação do modelo pode ser determinada plotando os sucessivos desvios em relação aos seus graus de liberdade e aceitando como uma medida do fator de escala a parte linear através da origem, determinada pelos pontos com menor número de graus de liberdade.

- Generalização da ANOVA

As primeiras diferenças dos desvios para a distribuição normal são (independente do fator de escala) a soma dos quadrados na análise de variância para um ajuste sequencial, como mostrado para um modelo de três termos na Tabela 1.

Tabela 1 - Desvio e suas diferenças

MODELO	DESVIO	DIFERENÇA	COMPONENTE
Mínimo	d_m	$d_m - d_A$	A
A	d_A	$d_A - d_{AB}$	B eliminando A
B	d_{AB}	$d_{AB} - d_{ABC}$	C eliminando A e B
C	d_{ABC}	$d_{ABC} - d_0$	Resíduo

Completo	d_0		

Fonte: Nelder e Wedderburn (1972).

A análise de variância generalizada para um modelo sequencial é agora definida para ter componentes representados pelas primeiras diferenças do desvio, com graus de liberdade como definido acima. Esses componentes têm distribuições proporcionais para X^2, exatamente para erros normais e aproximadamente para os outros. Tal generalização da análise da variância foi sugerida por Good em 1967.

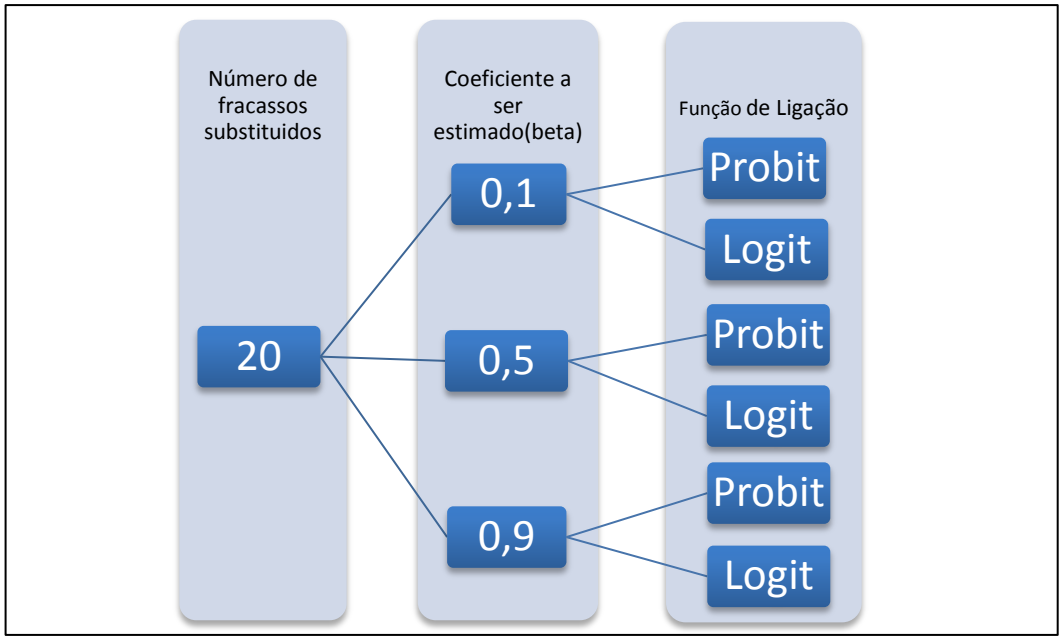
- Ajuste de modelos lineares generalizados no software R

Os MLG's são ajustados no R por meio da função ***glm*** (*formula, family*), onde devemos especificar formula (definição do modelo) e family (a distribuição da variável resposta com a função de ligação a ser utilizada), como por exemplo, MLG=glm (y~1+x, family=gaussian). É valido lembrar que "*gaussian*" corresponde à distribuição Normal. Caso a função de ligação usada seja diferente do "default", basta informar a função de ligação desejada através do comando link. Por exemplo, MLG=glm(y~1+x, family=gaussian(link="log")). O comando *summary(MLG)* fornece um resumo do resultado do ajuste.

2 METODOLOGIA

Inicialmente foi criada uma variável “x” que segue a distribuição normal padrão com duzentas observações através do comando “*rnorm(200)”* para ser usada como variável independente na formulação do Modelo Linear Generalizado. Em seguida, foram definidas três valores para o parâmetro beta a ser utilizada conjuntamente com a variável “x” na criação do vetor de médias (mu) que foi utilizado na criação da variável “y” que segue a distribuição Bernoulli também com duzentas observações através do comando “*rbinom(200, size=1, prob=mu)*” para ser usada como variável dependente no MLG. Abaixo segue o fluxograma com base nos casos simulados.

Figura 1 – Fluxograma do Método



Fonte: Elaboração própria (2015).

Então foram realizadas 1000 simulações para cada combinação de parâmetro e função de ligação. Em cada simulação foi alterado 10% (20) dos fracassos da amostra gerada em sucessos, o que eventualmente aumentará a estimativa da probabilidade de sucessos. Procurou-se estimar esse parâmetro conhecido através das funções de ligação logit e probit e observar qual das duas funções de ligação apresentaria a melhor estimação de beta, no nosso caso, qual das duas funções de ligação produziria o menor Erro Quadrático Médio. Todo o procedimento de análise foi realizado através do Software livre R (ver Apêndice).

3 RESULTADOS E CONCLUSÕES

Para o valor de beta igual a 0,1 após as 1000 simulações foi obtido o valor do Erro Quadrático Médio, que para o caso em que foi utilizada a função de ligação logit correspondeu a 0,021, já quando a função de ligação probit foi utilizada gerou-se um erro quadrático médio menor cujo valor foi 0,0099, o que nos leva a concluir que a função de ligação probit se mostrou mais estável ao erro gerado com a substituição dos fracassos por sucessos na distribuição da variável y. Sendo assim para o caso de beta igual a 0,1 temos a função de ligação probit como aquela que mais se adéqua a distribuição Bernoulli (Tabela 2).

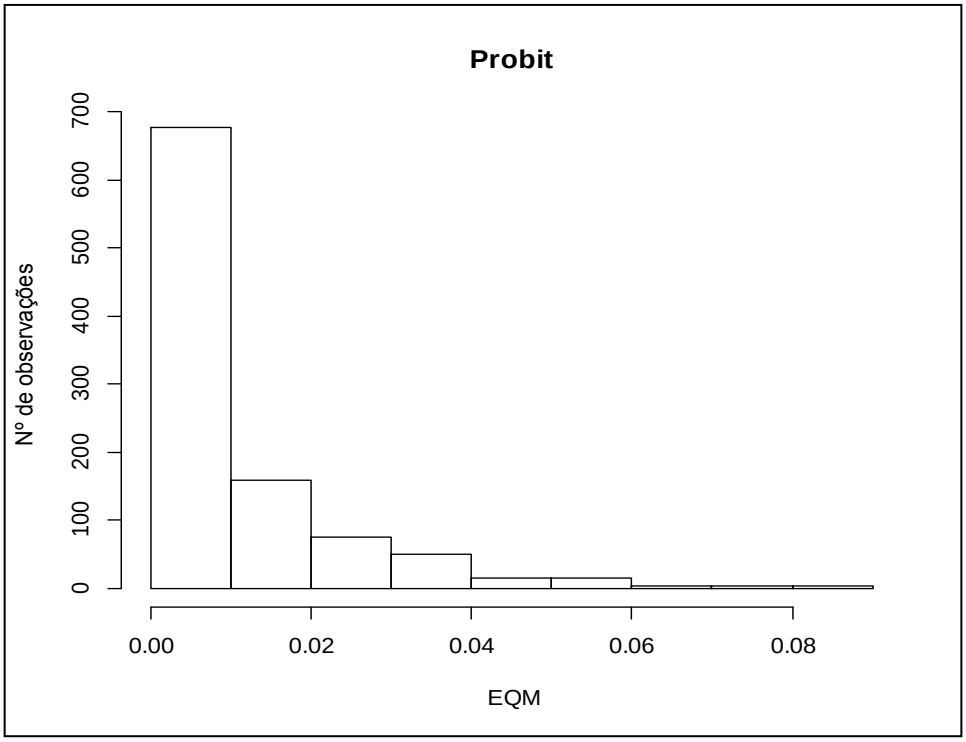
Tabela 2 – Valores para comparação das funções (beta=0,1)

COEFICIENTES\FUNÇÕES	LOGIT	PROBIT
EQM	0,02189647	0,009957982
Porcentagem de Erros <= 0,02	66,1	83,5

Fonte: Elaboração própria (2015).

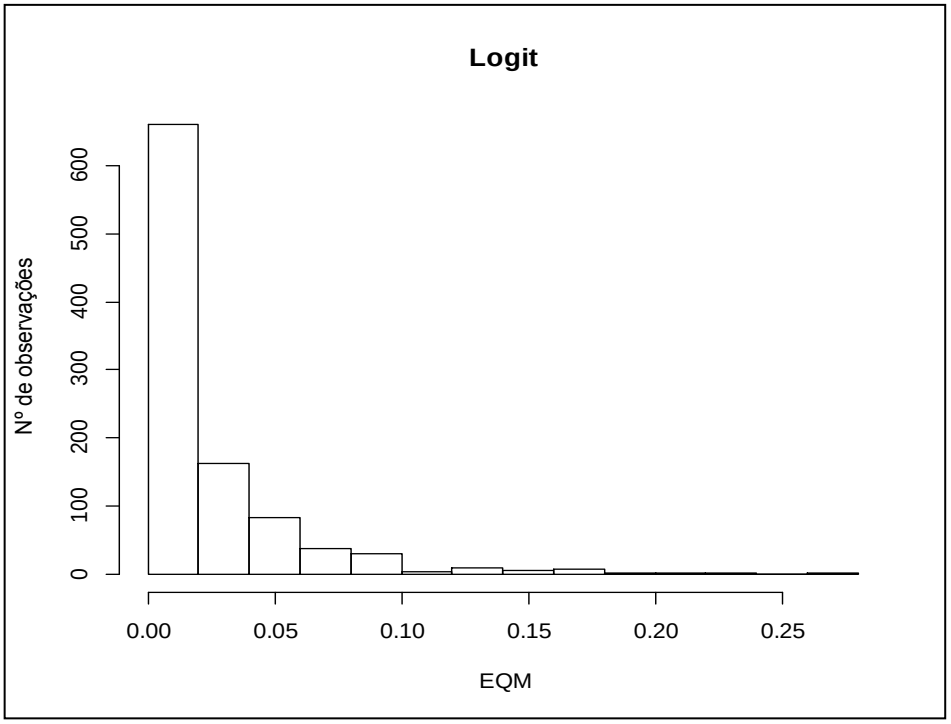
É valida ainda para efeito de comparação a informação de que das 1000 simulações realizadas em 66,1% delas o erro quadrático foi inferior a 0,02 para o logit e 83,5% para o probit.

Figura 2 – Histograma da Distribuição do EQM para a função de ligação Probit (beta=0,1)



Fonte: Elaboração própria (2015).

Figura 3 – Histograma da Distribuição do EQM para a função de ligação Logit (beta=0,1)



Fonte: Elaboração própria (2015).

Para o valor de beta igual a 0,5 após as 1000 simulações foi obtido o valor do Erro Quadrático Médio, que para o caso em que foi utilizada a função de ligação logit correspondeu a 0,033. Já quando a função de ligação probit foi utilizada gerou-se um erro quadrático médio menor cujo valor foi 0,014, o que nos leva a concluir que a função de ligação probit se mostrou novamente mais estável ao erro gerado com a substituição dos fracassos por sucessos na distribuição da variável y. Sendo assim para o caso de beta igual a 0,5 temos a função de ligação probit como aquela que melhor se adapta a distribuição Bernoulli (Tabela 3).

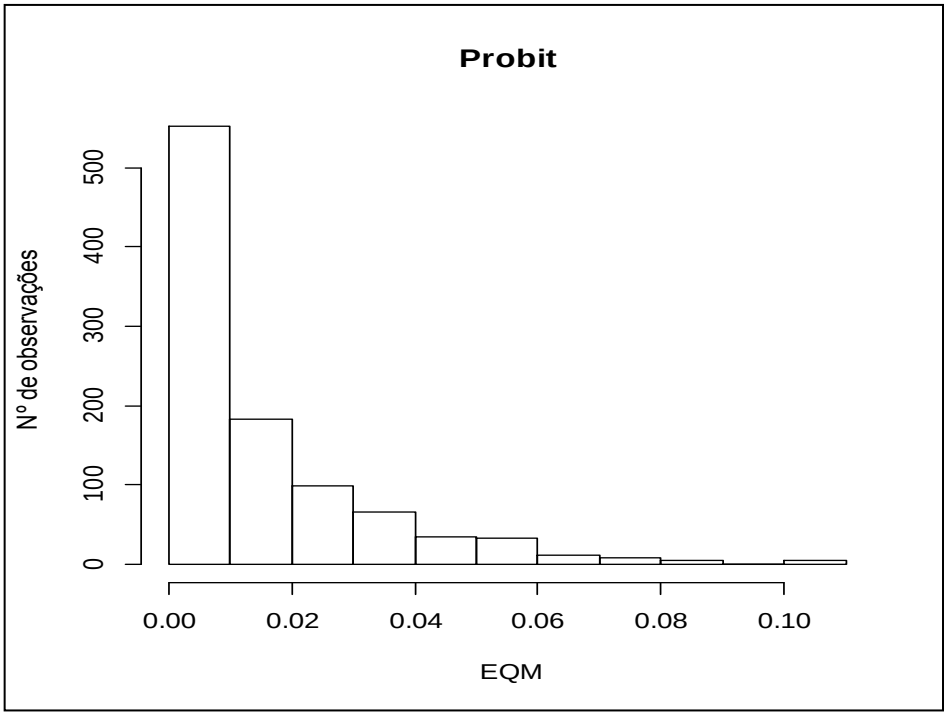
Tabela 3 – Valores para comparação das funções (beta=0,5)

COEFICIENTES\FUNÇÕES	LOGIT	PROBIT
EQM	0,03378041	0,01486734
Porcentagem de Erros <= 0,02	55,1	73,6

Fonte: Elaboração própria (2015).

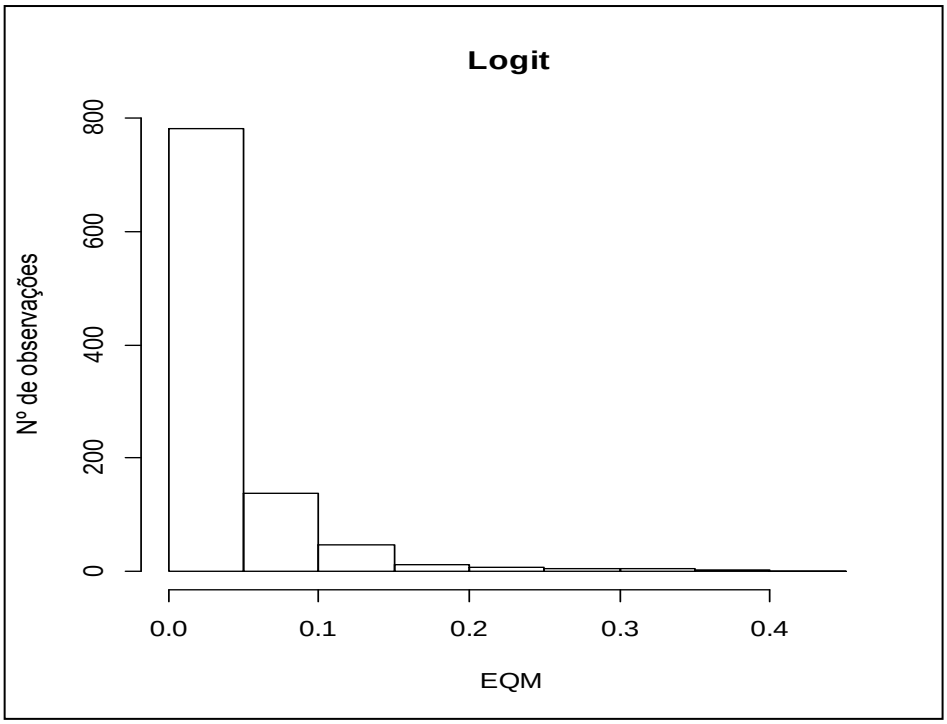
É valida ainda para efeito de comparação a informação de que das 1000 simulações realizadas em 55,1% delas o erro quadrático foi inferior a 0,02 para o logit e 73,6% para o probit.

Figura 4 – Histograma da Distribuição do EQM para a função de ligação Probit (beta=0,5)



Fonte: Elaboração própria (2015).

Figura 5 – Histograma da Distribuição do EQM para a função de ligação Logit (beta=0,5)



Fonte: Elaboração própria (2015).

Para o valor de beta igual a 0,9 após as 1000 simulações foi obtido o valor do Erro Quadrático Médio, que para o caso em que foi utilizada a função de ligação logit correspondeu a 0,05643825. Já quando a função de ligação probit foi utilizada gerou-se um erro quadrático médio menor cujo valor foi 0,1151118, o que nos leva a concluir que dessa vez a função de ligação logit se mostrou mais estável ao erro gerado com a substituição dos fracassos por sucessos na distribuição da variável y. Sendo assim para o caso de beta igual a 0,9 temos a função de ligação logit como aquela que melhor se adapta a distribuição Bernoulli (Tabela 4).

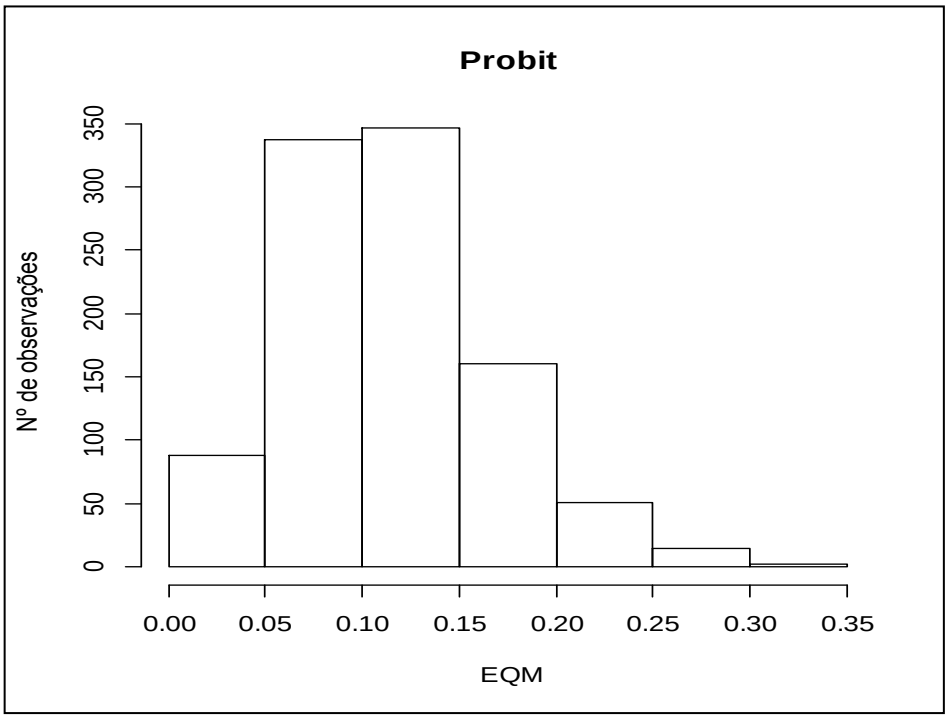
Tabela 4 – Valores para comparação das funções (beta=0,9)

COEFICIENTES\FUNÇÕES	LOGIT	PROBIT
EQM	0,05643825	0,1151118
Porcentagem de Erros <= 0,02	41,1	0,9

Fonte: Elaboração própria (2015).

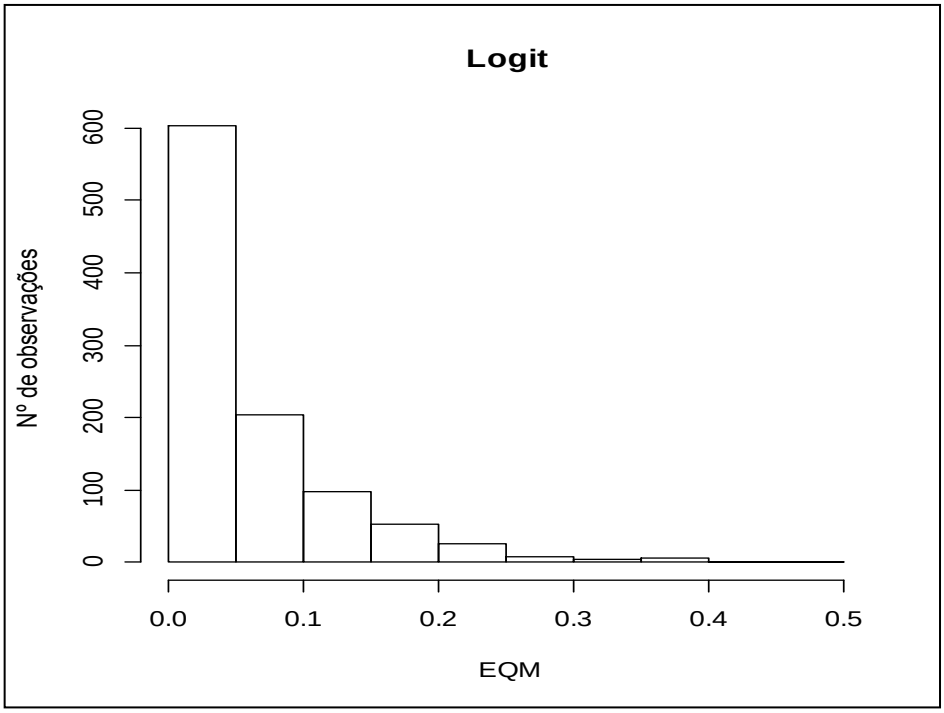
É valida ainda para efeito de comparação a informação de que das 1000 simulações realizadas em 41,1% delas o erro quadrático foi inferior a 0,02 para o logit e apenas 0,9% para o probit.

Figura 6 – Histograma da Distribuição do EQM para a função de ligação Probit (beta=0,9)



Fonte: Elaboração própria (2015).

Figura 7 – Histograma da Distribuição do EQM para a função de ligação Logit (beta=0,9)



Fonte: Elaboração própria (2015).

Após as análises realizadas chegou-se a conclusão que a função de ligação probit mostrou-se a melhor opção para estimação quando o parâmetro beta assumiu valores próximos a zero e a função de ligação logit foi a mais indicada para estimar quando o parâmetro beta assumiu valores próximos a um.

4 PERSPECTIVAS DE TRABALHOS FUTUROS

Este trabalho abre uma discussão sobre a adequação das funções de ligação aos dados, não considerando somente os tipos de dados das distribuições, mas também os valores dos parâmetros das mesmas. Com um maior prazo de execução seria possível realizar outros estudos para verificar a utilidade das diversas funções de ligação para as varias distribuições dos dados, com relação aos valores dos parâmetros das distribuições.

Também seria interessante verificar se a variação do número de observações substituídas (no nosso caso 10%) teria alguma influência na determinação de qual seria a melhor função de ligação para cada caso. De mesma maneira pode-se pensar em alterar sucessos para fracassos e diminuir a estimativa da probabilidade de sucesso para ver o impacto neste trabalho.

REFERÊNCIAS

ASEVEDO, F. R. **Abordagem linear generalizada para estimar perdas não técnicas de energia elétrica**. 2011. Dissertação (Mestrado em Engenharia Elétrica)–Pontifícia Universidade Católica Do Rio De Janeiro - PUC-RIO, Rio de Janeiro, 2011.

CADIMA, J. **Modelos Lineares Generalizados**. Universidade de Lisboa, 2010. Disponível em: <http://www.isa.utl.pt/dm/mestrado/2009-10/UCs/me2/slidesGLM.pdf>. Acesso em: 17 jan. 2015.

CORDEIRO, G.M. **Modelos Lineares Generalizados**. Campinas, VII SINAPE, 1986. 286p.

CORDEIRO, G.; DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados e Extensões**. Universidade de São Paulo - USP, 2010. Disponível em:< http://www.lce.esalq.usp.br/arquivos/aulas/2010/LCE5868/livro.pdf>. Acesso em: 17 jan. 2015.

CYSNEIROS, F. J. A. **Aula – Modelos Lineares Generalizados**. Universidade Federal de Pernambuco. Disponível em: <http://www.de.ufpe.br/~cysneiros/disciplina/MES940/aulaMLGmestrado.pdf>. Acesso em: 03 nov. 2014.

DEMÉTRIO, C. G. P. **Modelos Lineares Generalizados em Experimentação Agronômica.** Universidade de São Paulo - USP, 2002. Disponível em: <http://www.lce.esalq.usp.br/clarice/Apostila.pdf>. Acesso em: 17 jan. 2015.

FREITAS, L. R. et al. Comparação das funções de ligação logit e probit em regressão binária considerando diferentes tamanhos amostrais. **Enciclopédia Biosfera**, Goiânia, v.9, n.17; p. 2936-2951, dez. 2013.

LATORRE, M. R. D. O. **Lista de funções do R**. Universidade de São Paulo - USP, 2013. Disponível em:< www.fsp.usp.br/~rosario/r/listacodigos.pdf>. Acesso em: 16 jan. 2015.

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

PAULA, G. A. **Modelos de Regressão com apoio computacional**. Universidade de São Paulo, 2013. Disponível em: <http://www.ime.usp.br/~giapaula/texto_2013.pdf>. Acesso em: 17 jan. 2015.

PROVETE, D. B.; SILVA, F. R.; SOUZA, T. G. **Estatística aplicada à ecologia usando o R**. Universidade Estadual Paulista - UNESP, 2011. Disponível em: <http://cran.r-project.org/doc/contrib/Provete-Estatistica_aplicada.pdf>. Acesso em: 17 jan. 2015.

RESENDE, M. D. V.; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binominais. **Revista de Matemática e Estatística**, São Paulo, v.20; p. 39-65, 2002.

SOUZA, H. S. E.; LEÃO, L. C. S. Tarifação de um plano de saúde autogestão aplicando os Modelos Lineares Generalizados. **CADERNOS DO IME – Série Estatística**, Rio de Janeiro, v.33; p. 01-17, 2012.

TURKMAN, M. A. A.; SILVA, G. L. **Modelos Lineares Generalizados - da teoria à prática**. Universidade de Lisboa, 2000. Disponível em: <http://docentes.deio.fc.ul.pt/maturkman/mlg.pdf>. Acesso em: 17 jan. 2015.

VENABLES, W. N.; SMITH, D. M. **An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.1.2 (2014-10-31)**. R Core Team, 2014. Disponível em: <http://cran.r-project.org/doc/manuals/R-intro.pdf>. Acesso em: 17 jan. 2015.

VIEIRA, A. F. C. **Análise da média e dispersão em experimentos fatoriais não replicados para otimização de processos industriais**. 2004. Tese (Doutorado em Engenharia de Produção)–Pontifícia Universidade Católica Do Rio De Janeiro - PUC-RIO, Rio de Janeiro, 2004.

APÊNDICE – Comandos do Software R

```
#################################################################
## Calculando o Erro Quadrático Médio p/ os diversos valores de beta ##
#################################################################
i=1
eqmlogit=0
repeat
{
## valores da covariável:
x <- rnorm(200);n <- length(x)
## Definindo o valor do parâmetro $\beta$
beta=0.1 #alterar para outros valores de beta (0.5 e 0.9)
## Modelo Binomial (Bernoulli) com ligação "logit"
## calculando o vetor de médias $\mu$
mu <- exp(x * beta)/(1+ exp(x*beta))
## Simulando dados
y <- rbinom (n, size=1, prob=mu)
# alterando 20 fracassos em sucessos
alt=0
k=1
repeat
{
if (y[k]==0) { y[k]=1 ; alt=alt+1}
if (alt==20) break;
k=k+1
}
logit20=glm(y~x,family=binomial(link = "logit"))
eqmlogit[i]=(coef(logit20)[2]-beta)^2
if (i==1000) break;
i=i+1
}
## Modelo Binomial (Bernoulli) com ligação "probit"
i=1
eqmprobit=0
repeat
{
## calculando o vetor de médias $\mu$
mu <- pnorm(x*beta)
y <- rbinom(n, size=1, prob=mu)
# alterando 20 fracassos em sucessos
alt=0
k=1
repeat
{
if (y[k]==0) { y[k]=1 ; alt=alt+1}
if (alt==20) break;
k=k+1
}
```

```
probit20=glm(y~x,family=binomial(link = "probit"))
eqmprobit[i]=(coef(probit20)[2]-beta)^2
if (i==1000) break;
i=i+1
}
mean(eqmlogit)
mean(eqmprobit)
#calculando porcentagem de Erros Quadráticos <=0,2
alt=0
k=1
soma1=0
soma2=0
repeat
{
if (eqmprobit[k]<=0.02) { soma1=soma1+1 }
if (eqmlogit[k]<=0.02) { soma2=soma2+1 }
if (alt==1000) break;
k=k+1
alt=alt+1
}
soma1/1000
soma2/1000
par(mfrow=c(1,2))
hist(eqmprobit, main='Probit', ylab='Nº de observações', xlab=' EQM')
hist(eqmlogit, main='Logit', ylab='Nº de observações', xlab=' EQM')
```