COMPARAÇÃO DAS FUNÇÕES DE LIGAÇÃO LOGIT E PROBIT EM REGRESSÃO BINÁRIA CONSIDERANDO DIFERENTES TAMANHOS AMOSTRAIS

Leillimar dos Reis Freitas[1], Sebastião Martins Filho[2], José Ivo Ribeiro Júnior[2], Fabyano Fonseca e Silva[2]

1. Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa
2. Professor do Departamento de Estatística da Universidade Federal de Viçosa, martinsfilho@ufv.br, UFV, Viçosa-MG – Brasil

Recebido em: 30/09/2013 – Aprovado em: 08/11/2013 – Publicado em: 01/12/2013

RESUMO

Neste estudo foi considerada a regressão para resposta binária por meio das funções de ligação logit e probit, com finalidade de verificar a robustez destas funções diante da variação do tamanho da amostra. Foram realizadas simulações com 500 repetições utilizando amostras de 10 diferentes tamanhos, desde 10 a 91, com uma diferença de 9 unidades entre as amostras sucessivas. As medidas de desempenho: percentual de convergência, erro quadrático médio da probabilidade geral, erro quadrático médio da probabilidade específica e teste Wald para os coeficientes foram utilizadas para estabelecer uma avaliação do uso das duas funções de ligação quando os dados foram gerados com o uso do logit e probit e analisados por ambas as funções de ligação, para cada tamanho de amostra. Foi possível verificar relevantes diferenças entre as regressões, ao estabelecer o uso da função de ligação logit para tamanhos inferiores a 20, devido a maior taxa de convergência. Para tamanhos de amostras maiores utilizando as demais medidas de desempenho, tanto o logit como o probit mostraram-se semelhantes.
PALAVRAS-CHAVE: Simulação. Tamanho de amostra. Variável dicotômica

COMPARISON OF LOGIT AND PROBIT LINK FUNCTIONS IN BINARY REGRESSION CONSIDERING DIFFERENT SAMPLE SIZES

ABSTRACT

It was considered a binary regression analysis with the logit and probit link function in order to verify the link functions robustness in sample size variation. Were performed simulations with 500 replicates using 10 different sizes samples, from 10 to 91, with 9 successively units between the samples. Performance convergence percentage measures, general probabilistic average squared error, specific probabilistic average squared error and coefficients Wald test were used to establish a specific use recommendation for the two different link functions just when data were generated with logit and probit use and analyzed with the both link functions, in each sample size. It was possible to verify relevant differences between the regressions to establish the use of the logit link for sizes below 20, due to higher rates of convergence. For larger sample sizes, using other measures of performance, both the logit and the probit were similar.
KEYWORDS: Dichotomous variable. Sample size. Simulation

INTRODUÇÃO

Em muitas aplicações do modelo de regressão linear pressupõe-se que a variável dependente é uma variável aleatória contínua com distribuição normal para os erros. Existe, no entanto, situações em que a variável dependente não tem natureza contínua, como é o caso de variáveis discretas. Neste caso, o domínio da variável é o conjunto dos números inteiros e a hipótese de normalidade não é adequada. Para essa classe de variáveis dispõe-se de muitos modelos que podem ser utilizados, particularmente os chamados modelos de escolha binária em que a escolha faz-se entre duas alternativas e uma, ou outra, tem de ser escolhida. (CADIMA, 2013).

Os modelos lineares generalizados (MLGs) são uma família muito vasta de modelos que generalizam o modelo linear. A generalização dos MLGs incide essencialmente sobre dois aspectos fundamentais: a distribuição de probabilidades associada à variável-resposta aleatória Y já não se restringe à Normal, podendo ser qualquer distribuição numa classe designada família exponencial de distribuições; a relação entre a combinação linear das variáveis preditoras e a variável dependente pode ser mais geral do que no modelo linear. (CADIMA, 2013).

Muitos modelos são casos especiais de modelos lineares generalizados que são compostos de três componentes: um componente aleatório que identifica a distribuição de probabilidades da variável dependente; um componente sistemático que consiste numa combinação linear de variáveis preditoras; e por uma função de ligação que é uma função diferenciável e monótona que associa as componentes aleatória e sistemática (McCULLAGH & NELDER, 1989).

Para BENDER FILHO et al. (2010), uma maneira adequada de ajustar um modelo para variáveis binárias é pelas probabilidades. Desse modo existem funções de ligações específicas como logit e probit, que com a utilização de funções de distribuições podem realizar a análise. De acordo com a abordagem realizada por CORDEIRO & DEMÉTRIO (2007), a função de ligação logit assim como a probit têm em comum o fato de a variável dependente ser uma variável qualitativa binária. Desta forma, as funções de ligação logit e probit são dadas respectivamente pelos inversos das distribuições acumuladas logística e normal.

A função de ligação mais difundida na área de melhoramento animal é a probit, conhecida como modelo *threshold* (GIANOLA & FOULLEY, 1983; ABDEL-AZIM e BERGER, 1999; KADARMIDEEN et al., 2000, 2001; SILVA et al., 2005; SHIOTSUKI et al., 2009), porém existem outras funções que podem ser exploradas, como a logit, que é amplamente utilizada na área de biometria (BRESLOW & CLAYTON, 1993; DEMÉTRIO, 2001; NUNES et al., 2004).

De acordo com BARROS (2008) a escolha da função de ligação logit ou probit é determinada por simples conveniência matemática e computacional, no entanto, devido à diferença nas formas das curvas representativas destas distribuições é importante avaliar situações nas quais uma ou outra descrevem com maior precisão a probabilidade de interesse.

O presente trabalho teve como objetivo verificar o efeito do tamanho da amostra sobre a qualidade de ajuste e da robustez das funções de ligação logit e probit no ajuste da regressão de uma variável dependente dicotômica.

MATERIAL E MÉTODOS

SIMULAÇÃO DE DADOS

Para a realização da simulação, inicialmente foram definidos o tamanho da amostra, o tipo de equação utilizada (quantidade de variáveis dependentes e parâmetros), os valores correspondentes da variável independente e os parâmetros da equação a ser utilizada.

O valor assumido para a variável independente (*x*) foi definido pela divisão do intervalo de 1 a 10 em 10 diferentes valores (10, 20, 30, 40, 50, 60, 70, 80, 90, 100), assim obteve-se 10 diferentes tamanhos de amostra (*n*), conforme pode ser observado na Tabela 1.

TABELA 1 - Tamanhos das amostras iniciais, sequências da variável independente e novos tamanhos das amostras

Divisão do intervalo (1≤ *x* ≤ 10)	*X*	Tamanhos das amostras (*n*)
10	1,0; 2,0; 3,0; ... ; 10	10
20	1,0; 1,5; 2,0; ... ; 10	19
30	1,0; 1,33; 1,67; 2,0; ... ; 10	28
40	1,0; 1,25; 1,5; ... ; 10	37
50	1,0; 1,2; 1,4; 1,6; ... ; 10	46
60	1,0; 1,167; 1,333; ... ; 10	55
70	1,0; 1,14; 1,28; ... ; 10	64
80	1,0; 1,125; 1,250; ... ; 10	73
90	1,0; 1,11; 1,22; ...; 10	82
100	1,0; 1,10; 1,20; 1,30; ... ; 10	91

Os tamanhos de amostras foram determinados de forma que em tamanhos pequenos, se espera a maior ocorrência de erros, e tamanhos maiores em que há diminuição desta mesma estatística.

A equação considerada como referência para a realização do ajuste obtido utilizando as funções de ligação logit e probit foi definida somente com dois parâmetros:

$$\begin{cases} logit_i = g(p_i) = \beta_0 + \beta_1 x_i \\ probit_i = g(p_i) = \beta_0 + \beta_1 x_i \end{cases}$$

Estas equações foram consideradas verdadeiras servindo de comparação com as equações estimadas por meio dos dados simulados. O $logit_i$ (logit verdadeiro) e $probit_i$ (probit verdadeiro) foram definidos de formas iguais cujos parâmetros foram fixados em: $\beta_0 = -5,5$ e $\beta_1 = 1$, para 1≤ *x* ≤ 10, em que β_0 é a constante e β_1 é o coeficiente da regressão.

Estes valores foram definidos de forma que, tanto para o logit como o probit os valores das probabilidades verdadeiras alcançassem valores próximos de zero (0,01098694 para o logit e 0,000003398 para o probit) e próximos de 1 (0,98901306 para o logit e 0,999996600 para o probit, respectivamente, para o menor e maior

valor de X). Portanto, mesmo partindo de valores iguais para o logit e probit, as probabilidades, como foram calculadas por meio de diferentes funções apresentaram resultados diferentes, sendo $P(Y=1|X=x_i) = p_i$, como mostrado na Tabela 2.

TABELA 2 - Probabilidades de ocorrências de $Y=1|X=x_i$ calculadas por meio das funções de ligação logit e probit

X	Logit	Probit
$x_1=1$	Y ~ Ber (0,01098694)	Y ~ Ber (0,000003398)
$x_2=2$	Y ~ Ber (0,02931223)	Y ~ Ber (0,000232629)
$x_3=3$	Y ~ Ber (0,07585818)	Y ~ Ber (0,006209665)
$x_4=4$	Y ~ Ber (0,18242552)	Y ~ Ber (0,066807200)
$x_5=5$	Y ~ Ber (0,37754067)	Y ~ Ber (0,308537500)
$x_6=6$	Y ~ Ber (0,62245933)	Y ~ Ber (0,691462500)
$x_7=7$	Y ~ Ber (0,81757448)	Y ~ Ber (0,933192800)
$x_8=8$	Y ~ Ber (0,92414182)	Y ~ Ber (0,993790300)
$x_9=9$	Y ~ Ber (0,97068777)	Y ~ Ber (0,999767400)
$x_{10}=10$	Y ~ Ber (0,98901306)	Y ~ Ber (0,999996600)

De posse dos valores verdadeiros do logit e probit, obtiveram-se as respectivas probabilidades de $Y=1|X=x_i$:

$$p_i = P(Y=1|\ X=x_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}\text{ , para } 1\leq x_i \leq 10$$

$$p_i = P(Y=1|\ X=x_i) = \Phi(\beta_0+\beta_1 x_i)\text{, para } 1\leq x_i \leq 10$$

A partir das probabilidades verdadeiras calculadas, foram realizadas 500 simulações, baseadas na distribuição de Bernoulli, para os valores de Y, que assumiram valores iguais a zero ou um, dentro de cada x_i, tendo-se:

$$Y \mid x_i \sim \text{Ber}\ (p_i)\text{, para } p_i=p_{Li} \text{ e } p_i = p_{Pi}\ .$$

em que p_{Li} e p_{Pi} correspondem, respectivamente, às probabilidades das funções de ligação logit e probit.

Para cada tamanho amostral (n) foram obtidos valores observados de Y decorrentes das distribuições de probabilidades das respectivas variáveis, modeladas pelas distribuições Logística e Normal, respectivamente. Isso implicou em obter um banco de dados influenciado por dois fatores: tamanho amostral e tipo de função de ligação (logit ou probit).

A simulação foi realizada no software livre R (R DEVELOPMENT CORE TEAM, 2013). De acordo com os valores simulados de Y, realizaram-se 500 análises de regressão binária, ou seja, 500 repetições (simulações); para os 10 diferentes tamanhos de n baseando-se nos 2 tipos de funções de ligação, obtendo um total de

10.000 análises.

Desse modo foram estabelecidas duas variáveis independentes: tamanho de amostra (n=10, 19, 28, ..., 91) e tipo de função de ligação (logit e probit), que foram responsáveis pela variação dos valores observados de *y*(0,1).

AJUSTE DAS EQUAÇÕES DE REGRESSÃO BINÁRIA

De posse dos valores de *Y*, foram realizadas análises de regressão binária a partir das funções de ligação logit e probit para ambos os casos simulados. Desta forma, as análises foram separadas em duas grandes classes. A primeira utilizando os valores de *Y* simulados a partir das probabilidades obtidas por meio de função de ligação logit e a segunda por meio das probabilidades da função de ligação probit.

Isto implicou que a análise de regressão binária realizada por meio da função logit utilizou de dados que deveriam ser analisados propriamente pela função de ligação na qual os dados tiveram origem, e também por meio do outro tipo de função de ligação (probit). O mesmo foi feito utilizando a função de ligação probit, conforme esquema apresentado na Figura 1.

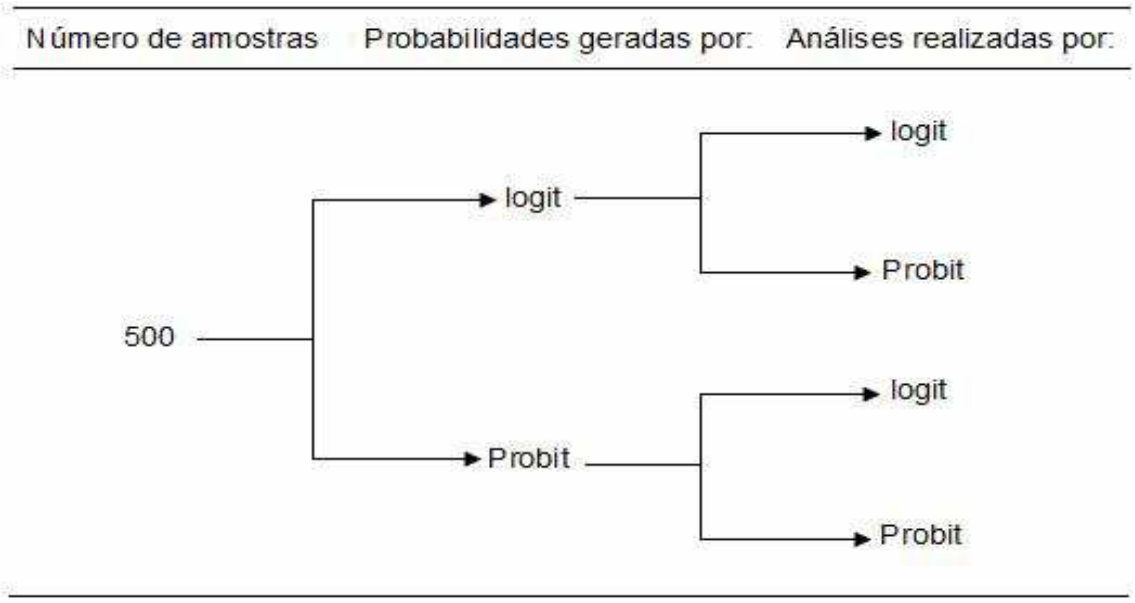

FIGURA 1 - Esquema das análises realizadas utilizando as funções de ligação logit e probit para cada tamanho de amostra.
Fonte: Autores

MEDIDAS DE DESEMPENHO

Após as obtenções das 500 equações de regressão binária, baseadas nas funções de ligação logit e probit, para cada valor de *n*, foram calculadas as seguintes medidas de desempenho:

i) Percentual de convergência: medida no qual determinado método iterativo se aproxima de seu resultado, ou seja, é o percentual das 500 equações binárias em que o algoritmo de Newton-Raphson se aproximou do verdadeiro valor (OLIVEIRA et al., 2000);

ii) Erro quadrático médio (LEHMAN, et al, 1998) da probabilidade geral estimada em relação à verdadeira: o cálculo dessa estatística foi obtido com a utilização de todos os diferentes valores de *x* (1 ≤ *x* ≤ 10), ou seja,

$$EQM\left[\hat{p}\right]=\frac{\sum_{i=1}^{n}\sum_{j=1}^{500}\left(\hat{p}_{ij}-p_{ij}\right)^2}{500n}$$

em que *n*=10,19,28,...,91;

iii) Erro quadrático médio (LEHMAN, et al, 1998) da probabilidade específica estimada em relação à verdadeira: seu cálculo foi obtido com a utilização dos níveis específicos de 1 ≤ *x* ≤ 10, ou seja, *x* iguais a 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10,

$$EQM\left[\hat{p}\right]_{x_1}=\frac{\sum_{j=1}^{500}\left(\hat{p}_{ij}-p_{ij}\right)^2}{500},$$

iv) Teste de Wald (WALD, 1943) dos parâmetros: foi utilizado para verificar quais as porcentagens de β_0, β_1 que foram significativamente diferente de zero, e também para verificar qual a porcentagem em que foi observado a constante e o coeficiente (ambos na mesma equação - β_0/β_1); Assim, o teste verificou a significância das seguintes hipóteses:

$$\begin{cases}H_0:\beta_0=-5,5\\H_1:\beta_0\neq-5,5\end{cases}\qquad\qquad\begin{cases}H_0:\beta_1=1\\H_1:\beta_1\neq1\end{cases}$$

Após a obtenção dos resultados de todas as medidas de desempenho utilizadas para a qualidade de ajuste das funções de ligação, foram realizadas análises de regressão destas em função do tamanho da amostra e do tipo de função de ligação de forma que para a realização da regressão o logit foi fixado como sendo 0 e o probit 1. Os coeficientes dos efeitos simples e de suas interações foram avaliados pelo teste t de Student (FISHER BOX, 1987) a 5% de probabilidade, ou seja, foi verificada a 5% a interação entre o tipo de função de ligação e o tamanho da amostra, a influência do tamanho da amostra e o tipo de função de ligação, isto é,

$$md_{\text{logit}}=\lambda_0+\lambda_1 n+\lambda_2 n^2+\lambda_3 f+\lambda_4 nf+\varepsilon,\quad \text{e}$$

$$md_{\text{probit}}=\gamma_0+\gamma_1 n+\gamma_2 n^2+\gamma_3 f+\gamma_4 nf+\varepsilon$$

em que *md* corresponde às medidas de desempenho obtidas pela regressão, λ_i e γ_i são parâmetros da equação, *n* o tamanho da amostra, e *f* o tipo de função de ligação que neste caso o logit assumiu o valor 0 e o probit 1.

RESULTADOS E DISCUSSÕES

PERCENTUAL DE CONVERGÊNCIA

O percentual de convergência do algoritmo (*c*) aumentou (*P-valor* < 0,05) somente em função do aumento de *n*, como segue,

c = -38,8778 + 7,17778* n, para 10 ≤ *n* < 19

c = 99,73, para 19 ≤ *n* ≤ 91

ou seja, o tamanho da amostra influenciou o percentual de convergência; a convergência também não é influenciada tanto pelo tipo de função de ligação quanto pela interação entre o tamanho da amostra e o tipo de função de ligação, em ambos os conjuntos de dados, o que também pode ser observado graficamente na Figura 2.

A convergência ocorreu em todos os casos quando o tamanho da amostra foi maior que 45 para os dois tipos de função de ligação (logit e probit). Para amostras menores que este tamanho, a convergência não ocorreu quando houve uma sequência gerada pelo *Y* do tipo em há uma sucessão de zeros seguidos por uns, ou seja, sequencias do tipo 0000011111, para *n*=10, tais resultados se referem aos valores de *X* iguais a 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10.

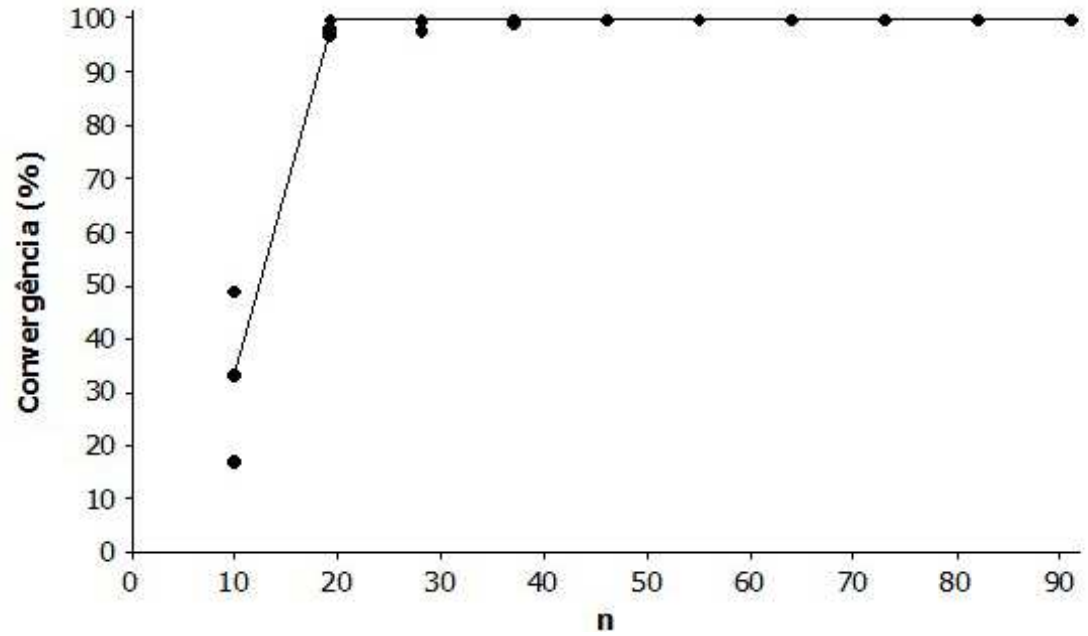

FIGURA 2 - Percentuais de convergências do logit e probit.
Fonte: Autores

De acordo com PENG et al. (2002) as estimativas dos coeficientes se tornam instáveis para pequenos tamanhos de amostras, o autor complementa que a literatura não oferece normas especificas quanto a determinação do tamanho que deva ser utilizado.

PEIXOTO et al. (2011) informa que a aplicação do modelo de regressão linear segmentada permite descrever o comportamento da variabilidade entre as variáveis, ou seja, a regressão segmentada foi utilizada pois permitiu descrever a variabilidade medida pelo percentual de convergência ao longo dos 10 diferentes tamanhos de amostras utilizados.

Portanto, quanto à convergência, tanto faz analisar os dados oriundos teoricamente de uma função de ligação logit ou probit, para amostras maiores que 20. Para amostras pequenas é recomendado o uso do logit devido à maior complexidade da função de ligação probit. Para as amostras em que o algoritmo

convergiu foi possível realizar as seguintes estatísticas.

ERRO QUADRÁTICO MÉDIO DA PROBABILIDADE GERAL

O erro quadrático médio diminuiu (P-valor < 0,05) em função do aumento de *n*, mais rapidamente para valores menores de *n* e tendendo a ser constante para os maiores valores. Ademais, não foi verificada diferença (P-valor > 0,05) entre as funções logit e probit. Os parâmetros nas equações (Figura 3 a e b) foram significativos pelo teste t de Student (P-valor < 0,05).

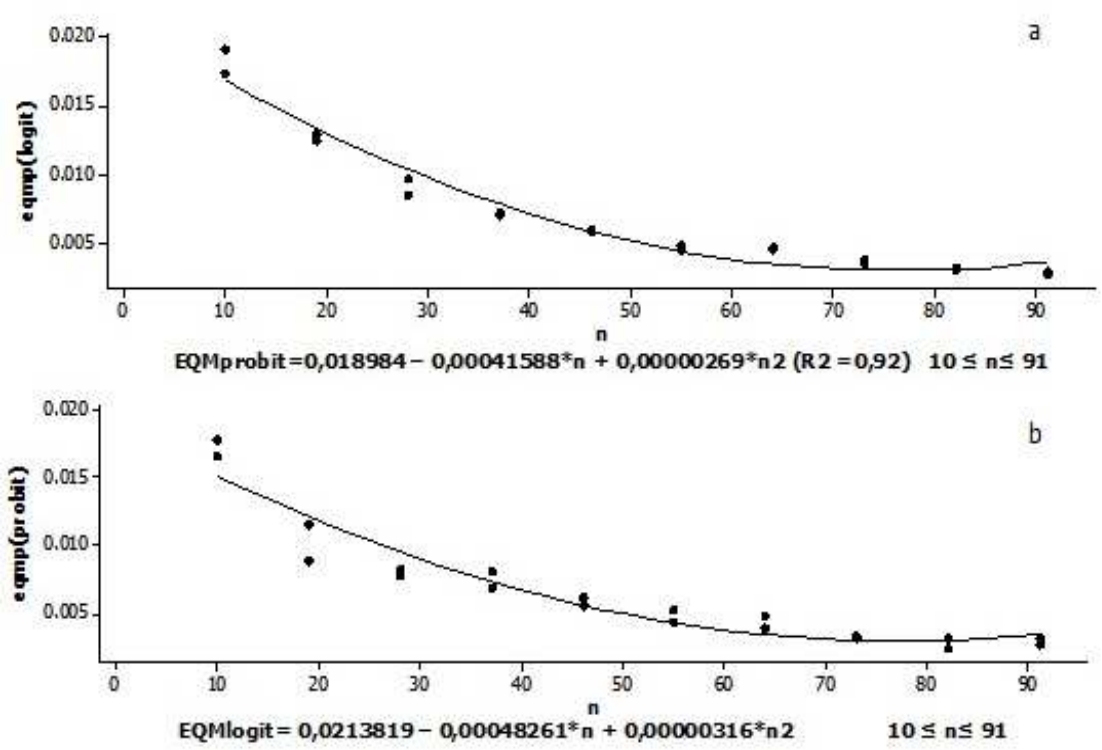

FIGURA 3 - Erro quadrático médio dos dados oriundos das funções de ligação logit e probit, em função do tamanho da amostra (*n*).
Fonte: Autores

Segundo MIOT (2011), o erro é inversamente proporcional ao tamanho da amostra, como pode ser observado na Figura 3, ou seja, à medida que o tamanho da amostra aumenta há uma diminuição do erro quadrático médio tanto do logit como do probit.

Como não foram observadas diferenças significativas entre as duas funções de ligação podem-se ajustar regressões binárias, tanto pela logit ou probit, ou seja, as duas funções de ligação possuem comportamento semelhante quanto ao erro quadrático médio em função do tamanho da amostra. Segundo O'DONNELL e CONNOR (1996), as estimativas de probabilidade do logit e probit são semelhantes. ESPAHBODI e ESPAHBODI (2003) reforça essa mesma teoria.

Recomenda-se que a amostra possua no mínimo 75 unidades, pois o erro quadrático médio diminui intensamente até esse tamanho de amostra.

De acordo com as duas equações de regressão foi verificado que se, teoricamente, a função é logit ou probit, podem-se estimá-las por meio das funções

logit ou probit, sem nenhum problema de ajuste. Isso implica que, a princípio, não é necessário conhecer qual é a melhor função para a obtenção do menor erro quadrático médio.

ERRO QUADRÁTICO MÉDIO DA PROBABILIDADE ESPECÍFICA

O erro quadrático médio fixados *x*=1,2, ... ,10 para as função de ligação logit e probit diminui em função do aumento de *n* (P-valor < 0,05). Além disso, não houve diferença entre os dois tipos de funções de ligação empregadas (P-valor > 0,05).

Para os dados simulados a partir da função de ligação logit, fixados diferentes níveis de *x,* como mostrados na Figura 4, obtiveram-se as equações de regressão ajustadas mostradas na Tabela 3.

TABELA 3 - Equação de regressão e grau de ajustamento quanto ao erro quadrático médio quadrático da probabilidade dos dados que foram originados da função de ligação logit fixados diferentes níveis de *x*

Níveis de *x*	Equação de Regressão*	R^2
x=1	$\hat{EQM}$ = 0,00420 - 0,000132* n + 0,000001*n^2	0,64
x=2	$\hat{EQM}$ = 0,00693 - 0,000216* n + 0,000002*n^2	0,60
x=3	$\hat{EQM}$ = 0,0196 - 0,000510* n + 0,000004* n^2	0,57
x=4	$\hat{EQM}$ = 0,0286 - 0,000590* n + 0,000004*n^2	0,84
x=5	$\hat{EQM}$ = 0,0412 - 0,000903* n + 0,000006* n^2	0,60
x=6	$\hat{EQM}$ = 0,0499 - 0,00109* n + 0,000007 *n^2	0,75
x=7	$\hat{EQM}$ = 0,0318 - 0,000826* n + 0,000006* n^2	0,62
x=8	$\hat{EQM}$ = 0,0151 - 0,000386* n + 0,000003* n^2	0,58
x=9	$\hat{EQM}$ = 0,00432 - 0,000105* n + 0,000001* n^2	0,76
x=10	$\hat{EQM}$ = 0,00981 - 0,000342* n + 0,000003* n^2	0,51

*Significativo pelo teste t de Student (P-valor < 0,05).

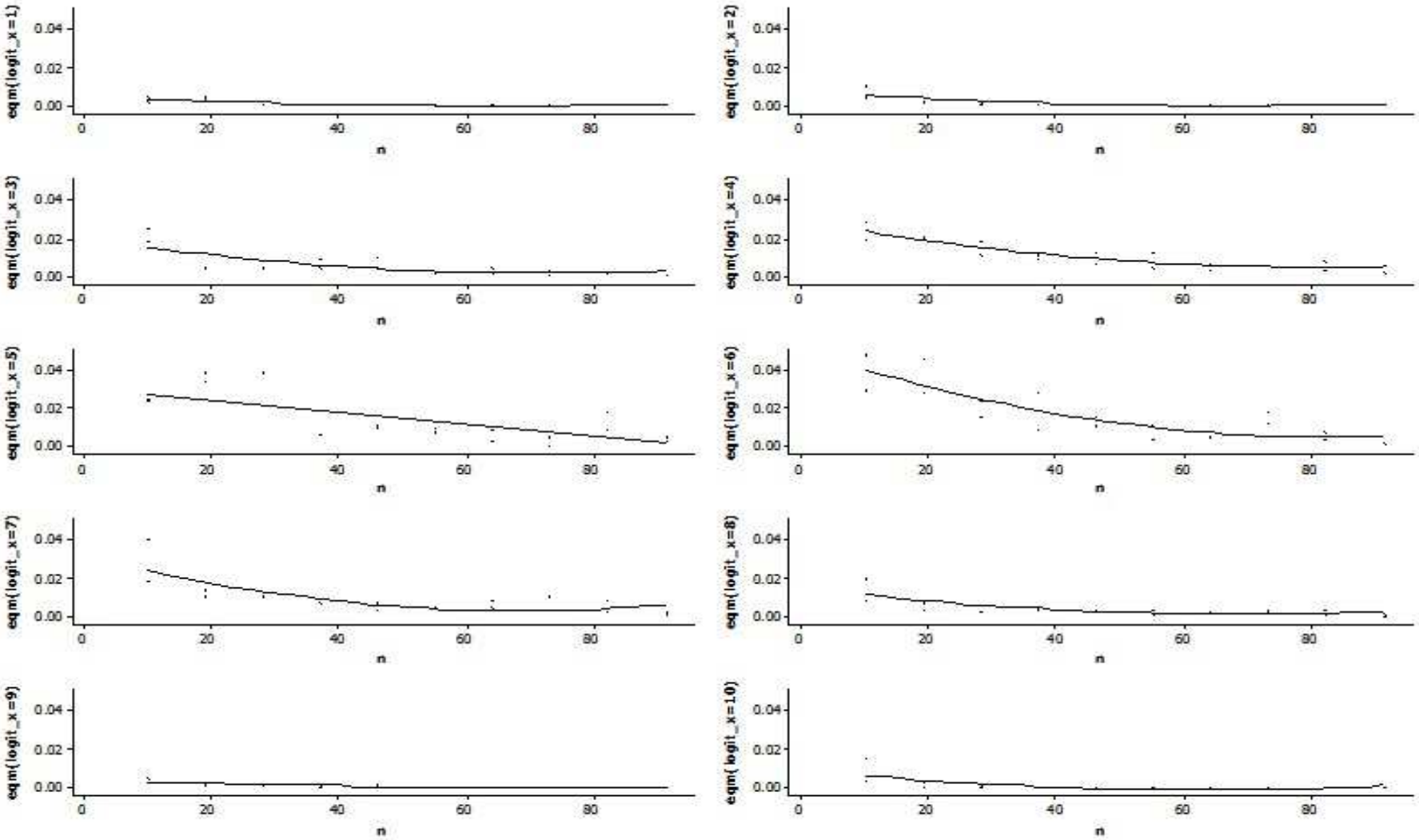

FIGURA 4 - Erro quadrático médio da probabilidade dos dados oriundos do logit fixados *x*=1, ...,10.
Fonte: Autores

Já para os dados simulados a partir da função de ligação probit fixados diferentes níveis de *x*, Tabela 4 e Figura 5, as equações de regressão ajustadas foram:

TABELA 4 - Equação de regressão e grau de ajustamento quanto ao erro quadrático médio quadrático da probabilidade dos dados que foram originados da função de ligação probit fixados diferentes níveis de *x*

Variável	Equação de Regressão*	R^2
x=1	$\hat{EQM}$ = 0,000134 - 0,000004 *n	0,29
x=2	$\hat{EQM}$ = 0,000835 - 0,000030 *n	0,39
x=3	$\hat{EQM}$ = 0,00715 - 0,000244* n + 0,000002* n^2	0,42
x=4	$\hat{EQM}$ = 0,0302 - 0,000961* n + 0,000008 *n^2	0,71
x=5	$\hat{EQM}$ = 0,0491 - 0,000833* n + 0,000005* n^2	0,80
x=6	$\hat{EQM}$ = 0,0451 - 0,000571* n + 0,000002* n^2	0,54
x=7	$\hat{EQM}$ = 0,0337 - 0,00106* n + 0,000008* n^2	0,44
x=8	$\hat{EQM}$ = 0,00564 - 0,000172* n + 0,000001* n^2	0,47
x=9	$\hat{EQM}$ = 0,000564 - 0,000018 *n	0,75
x=10	$\hat{EQM}$ = 0,000160 - 0,000006 *n	0,69

*Significativo pelo teste t de Student (P-valor < 0,05)

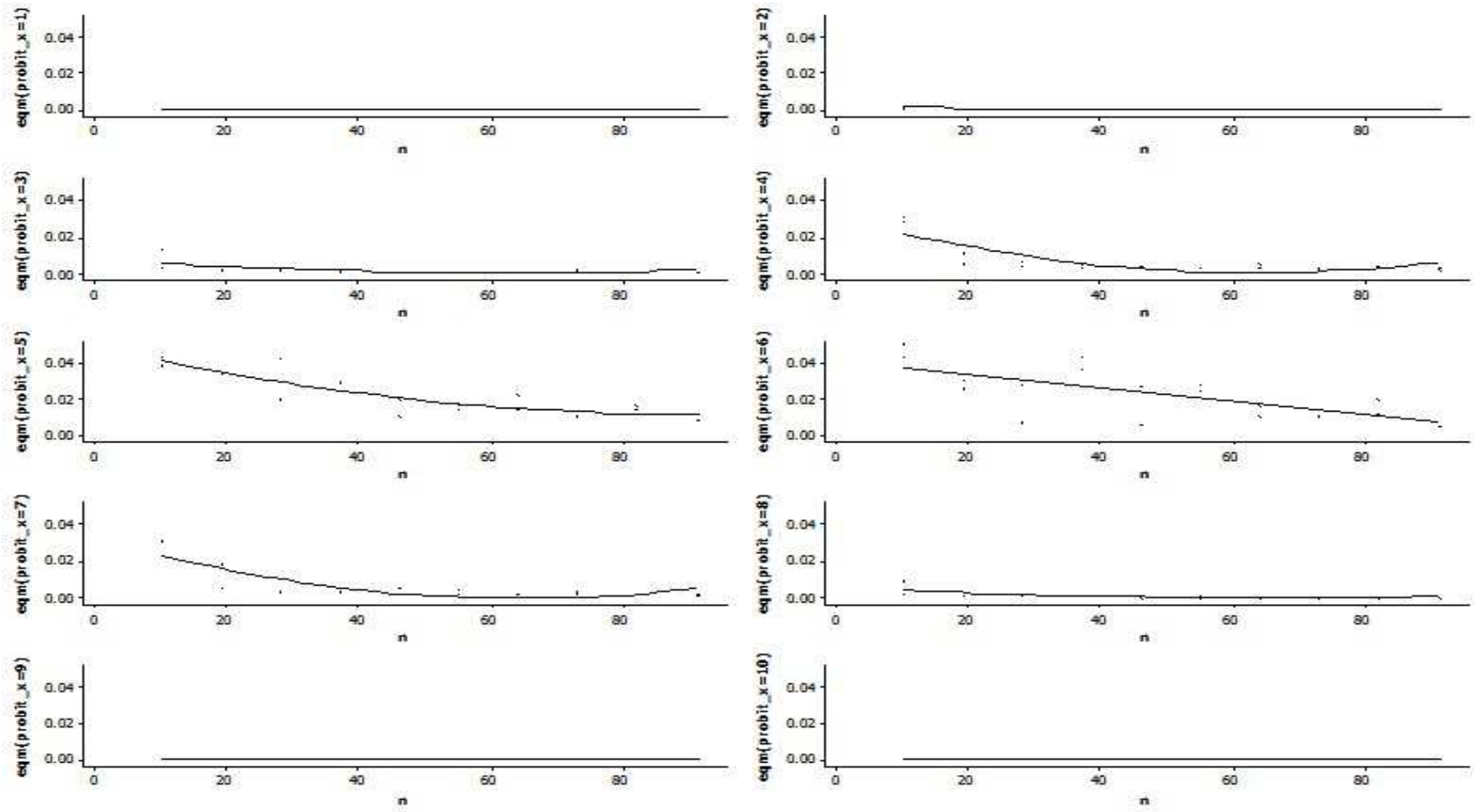

FIGURA 5 - Erro quadrático médio da probabilidade dos dados oriundos do probit fixados *x*=1, ..., 10.
Fonte: Autores

Verificou-se que para as funções de ligação logit e probit foram obtidas maiores estatística para o erro quadrático médio da probabilidade específica para valores intermediários de *x* (4, 5, 6, 7 e 8), enquanto que para valores extremos as elas foram menores, pois a diferença entre a probabilidade teórica e a estimada, quando há análise pelos dois tipos de funções de ligação é maior nesses valores intermediários, ou seja, as probabilidades nos extremos foram melhores estimadas.

De acordo com LONG (2009) as probabilidades previstas entre o logit e probit são quase idênticas, diferindo somente nas caudas devido ao tipo de distribuição utilizada para cada tipo de função de ligação. O autor complementa que tanto o logit como o probit o efeito de uma variável depende do nível de todas as outras variáveis.

TESTE DE WALD

Não houve diferença quanto à significância da estatística (*W*) do teste de Wald (P-valor > 0,05) dos dados que tiveram origem nas funções logit e probit como podem ser observados nas equações:

$W\beta_0\_\text{logit}$ = - 0,138 + 0,0189* n - 0,000115*n^2 R^2=0,98

$W\beta_1\_\text{logit}$ = - 0,133 + 0,0212 *n - 0,000138*n^2 R^2=0,97

$W\beta_0/\beta_1\_\text{logit}$ = - 0,135 + 0,0183* n - 0,000111 *n^2 R^2=0,98

$W\beta_0\_\text{probit}$ = - 0,163 + 0,0233* n - 0,000155*n^2 R^2=0,98

$W\beta_1\_\text{probit}$ = - 0,127 + 0,0244* n - 0,000169*n^2 R^2=0,94

$W\beta_0/\beta_1\_\text{probit}$ = - 0,162 + 0,0229*n - 0,000152*n^2 R^2=0,98

A significância dos parâmetros e da constante aumenta em função do aumento de *n* (P-valor < 0,05), mais rapidamente para valores iniciais menores que 60 e menos intensamente para os maiores valores de *n* até não mais exercer efeito (Figura 6).

De acordo com QUEIROZ (2011), o teste de Wald apresenta baixo desempenho em amostras pequenas; como pode ser observada na figura 6, a porcentagem de amostras em que os parâmetros foram significativos foi pequena para menores tamanhos de amostras. RAMALHO & RAMALHO (2009) complementa que o poder do teste Wald fica reduzido em pequenas amostras.

Portanto, podem-se ajustar regressões binárias, tanto pelas funções logit ou probit, recomenda-se no mínimo 60 pares de valores de *X* e *Y*. De acordo com as duas equações de regressão, foi verificado que se, teoricamente, a função é logit ou probit, podem-se estimá-las por meio das funções logit ou probit, sem nenhum problema de ajuste. Isso implica que, a priori, não é necessário conhecer qual é a melhor função.

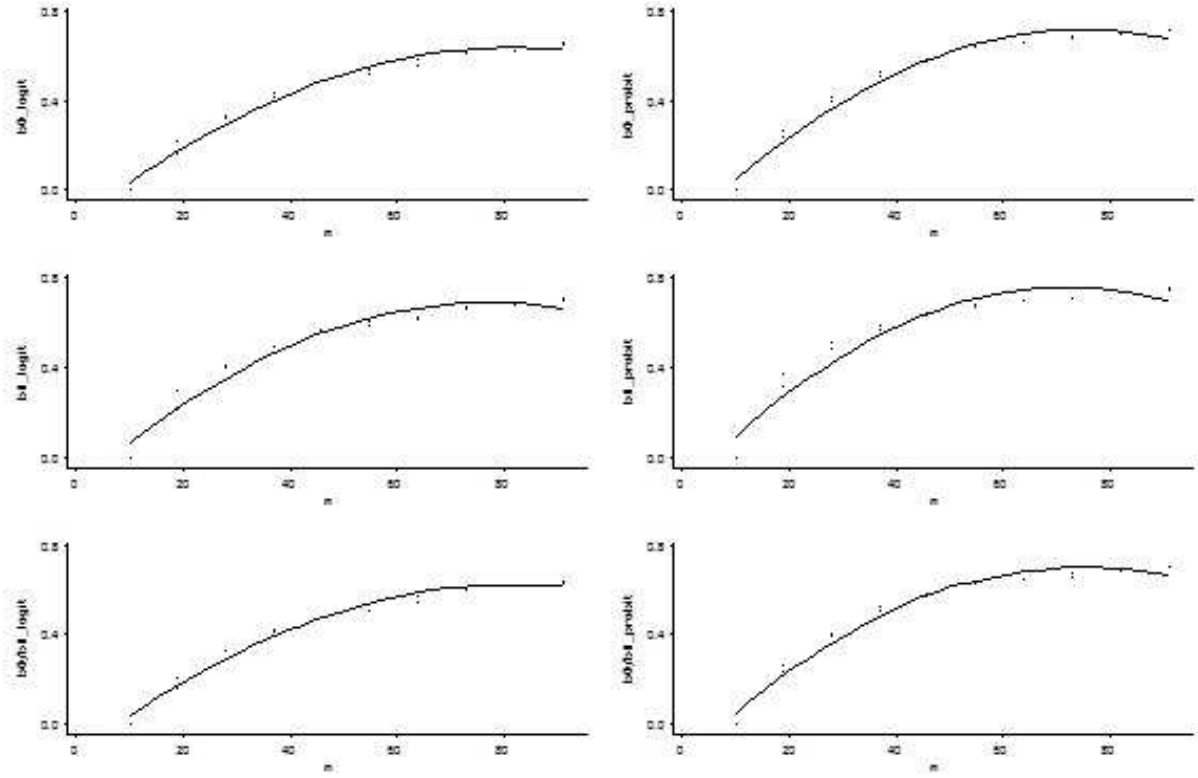

FIGURA 6 - Teste de Wald para a constante, coeficiente e constante com o coeficiente para os dados que tiveram origem nas funções de ligação logit e probit.

Fonte: Autores

CONCLUSÕES

A escolha da função pode ser subjetiva, mas o tamanho da amostra não, uma vez que ao aumentar o tamanho amostral melhora a qualidade do ajuste. Portanto, recomenda-se o uso da função de ligação logit para tamanhos inferiores a 20 e logit ou probit para maiores tamanhos de amostras.

AGRADECIMENTOS

A CAPES pela concessão da bolsa de L.R.Freitas.

REFERÊNCIAS

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

BARROS, G. C. O. **Modelos de previsão da falência de empresas: aplicação empírica ao caso das pequenas e médias empresas portuguesas**. (Dissertação) - Instituto Superior de Ciências do Trabalho e da Empresa -

Departamento De Economia - Lisboa, Portugal, 2008.

BENDER FILHO, R.; BAGOLIN, I, P.; COMIM, F. V. **Determinantes da permanência na condição de pobreza crônica: aplicação do modelo logit multinomial**. Texto para discussão. Porto Alegre. n. 07, 2010. Disponível em: <http://www3.pucrs.br/pucrs/ppgfiles/files/faceppg/ppge/texto_7_2010.pdf>. Acesso em: 22 set. 2013.

BRESLOW, N.E.; CLAYTON, D.G. Approximate inference in generalized linear mixed models. **J. Am. Stat. Assoc.**, v.88, p.9-25, 1993.

CADIMA, J. **Modelos lineares generalizados**. Lisboa:DM/ISA, 2013. Disponível em: <http://www.isa.utl.pt/dm/mestrado/2009-10/UCs/me2/slidesGLM.pdf>. Acesso em: 15 nov. 2013.

CORDEIRO, G.; DEMÉTRIO, C. Modelos lineares generalizados. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 12., e REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 52., 2007, Santa Maria. **Minicurso.** Santa Maria: UFSM, 2007. 161p.

DEMETRIO, C.G.B. **Modelos lineares generalizados em experimentação agronômica**. Piracicaba: ESALQ, 2001. 113p.

ESPAHBODI, H.; ESPAHBODI, P. Binary choice models and corporate takeover. **Journal of Banking & Finance**, v.27, p.549–574, 2003.

FISHER BOX, J. "Guinness, Gosset, Fisher, and Small Samples". **Statistical Science**, v.2, n.1, p. 45–52, 1987.

GIANOLA, D.; FOULLEY, J. L. Sire evaluation for ordered categorical data with threshold model. **Genet. Select. Evol.**, v.15, p.201-224, 1983.

KADARMIDEEN, H. N., THOMPSON, R.; SIMM, G. Linear and threshold model genetic parameters for disease, fertility and milk production in dairy cattle. **Anim. Sci.**, v.71, p.411-419, 2000.

KADARMIDEEN, H. N., REKAYA, R.; GIANOLA, D. Genetic parameters for clinical mastitis in Holstein-Friesians: A Bayesian analysis. **Anim. Sci.**, v.73, p.229-240, 2001.

LEHMANN, E. L.; CASELLA, G. **Theory of Point Estimation**. New York: Springer, 2nd, 1998.

LONG, J. S. **Group comparisons in logit and probit using predicted probabilities**. Indiana University, 2009

McCULLAGH, P.; NELDER, J.A. **Generalized Linear Models**. Chapman and Hall, Londres, 2nd, 1989.

MIOT, H. A. **Tamanho da amostra em estudos clínicos e experimentais**. Departamento de Dermatologia e Radioterapia da Faculdade de Medicina de Botucatu da Universidade Estadual Paulista (UNESP) - Botucatu (SP), Brasil, 2011.

NUNES, J. A. R.; MORAIS, A. R.; BUENO FILHO, J. S. S. Modelagem da superdispersão em dados por um modelo linear generalizado misto. **Rev. Mat. Estat.**, v.22, p.55-70, 2004.

O'DONNELL, C. J.; CONNOR, D. H. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. **Accident Analysis & Prevention**, v.28, n.6, p.739-753, 1996.

OLIVEIRA, H. N.; LÔBO, R. B.; PEREIRA, C. S.. Comparação de modelos não-lineares para descrever o crescimento de fêmeas da raça Guzerá. **Pesquisa Agropecuária Brasileira**, Brasília, v.35, n.9, p.1843-1851, 2000.

PEIXOTO, A. P.; FARIA, G. A.; MORAIS, A. R. Modelos de regressão com platô na estimativa do tamanho de parcelas em experimento de conservação in vitro de maracujazeiro. **Ciência Rural**, Santa Maria, v.41, n.11, p.1907-1913, 2011.

PENG, C. Y. J.; SO, T. S. H.; STAGE, F. K.; JOHN, E. P. S. The use and interpretation of logistic regression in higher education journals: 1988–1999. **Research in Higher Education**, v.43, p.259-293, 2002.

QUEIROZ, M. P. F. **Testes de hipóteses em regressão beta baseados em verossimilhança perfilada ajustada e em bootstrap**. 2011, 61 f. Dissertação (Mestrado em Estatística) - Universidade Federal de Pernambuco, Recife, 2011.

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, Version 3.0.1. Disponível em: http://www.R-project.org. Acesso em: 12 set. 2013.

RAMALHO, E. A.; RAMALHO, J. J. S. Is neglected heterogeneity really an issue in binary and fractional regression models? A simulation exercise for logit, probit and loglog models. Centro de estudo e formação avançada em gestão em economia - CEFAGE. **Working Paper**, n. 2009/10 - Universidade de Évora, Portugal, 2009.

SHIOSTSUKI, L.; SILVA, J. A. V.; ALBUQUERQUE, L.G. Associação genética da prenhez aos 16 meses com o peso à desmama e o ganho de peso em animais da raça Nelore. **Rev. Bras. Zootec.**, v.38, p.1211-1217, 2009.

SILVA, J.A.II V.; DIAS L.T.; ALBUQUERQUE L.G. Estudo genético da precocidade sexual de novilhas em um rebanho Nelore. **Rev. Bras. Zootec.**, v.34, p.1568-1572, 2005.

WALD, A. Test of statistical hypotheses concerning several parameters when the number of observation is large. **Trans. Amer. Math. Soc.**, v.54, p.426-482, 1943.