LEILLIMAR DOS REIS FREITAS

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

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

VIÇOSA
MINAS GERAIS – BRASIL
2013

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

T
F866c
2013

Freitas, Leillimar dos Reis, 1984-
Comparação das funções de ligação logit e probit em regressão binária considerando diferentes tamanhos amostrais / Leillimar dos Reis Freitas. – Viçosa, MG, 2013.
ix, 43f. : il. ; 29cm.

Inclui apêndice.
Orientador: Sebastião Martins Filho
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 33-35

1. Análise de regressão. 2. Métodos de simulação. 3. Amostragem (Estatística). I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.536

LEILLIMAR DOS REIS FREITAS

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

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

APROVADA: 20 de fevereiro de 2013.

Fabyano Fonseca e Silva
(Coorientador)

José Ivo Ribeiro Júnior
(Coorientador)

Fernanda Maria de Almeida

Sebastião Martins Filho
(Orientador)

DEDICATÓRIA

Minha família e
meu noivo Luís.

AGRADECIMENTOS

À Universidade Federal de Viçosa, pela oportunidade de realização deste curso.

A CAPES pelo apoio financeiro, esse trabalho só foi possível graças a bolsa que me foi concedida.

Aos professores do Departamento de Estatística, pelos ensinamentos, disponibilidade, amizade.

Ao professor Sebastião Martins Filho, pela orientação ao longo de todo o mestrado.

Ao professor Fabyano Fonseca e Silva pelo acompanhamento do meu trabalho, paciência, confiança e incentivo.

Ao professor José Ivo Ribeiro Júnior, pela coorientação no desenvolvimento deste trabalho.

Aos professores que participaram da banca examinadora por terem aceitado o convite e por suas contribuições oportunas, que certamente enriqueceram o trabalho.

Aos amigos de curso, pelo convívio agradável durante a realização deste curso.

Ao secretário do curso de pós-graduação em Estatística Aplicada e Biometria pelo apoio, dedicação, atenção e amizade.

Aos meus pais, Roseli e Joaquim, minhas irmãs, pela amizade, união e compreensão em todos os momentos.

Ao meu grande companheiro Luís, pela paciência e compreensão, obrigada por estar sempre me fazendo feliz e me apoiando em todos os momentos.

Aos amigos de outras datas que me apoiaram e me incentivaram em todos os momentos.

Por fim a todos que me ensinaram de alguma forma, me apoiaram, me acolheram e que fazem parte da minha história, muito obrigada a todos vocês.

BIOGRAFIA

LEILLIMAR DOS REIS FREITAS, filha de Joaquim Antônio de Freitas e Roseli dos Reis Freitas nasceu em Juiz de Fora, Minas Gerais, no dia 22 de outubro de 1984.

Graduou-se em Estatística pela Universidade Federal de Juiz de Fora em julho de 2010.

Em agosto de 2010 iniciou o mestrado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa (UFV) tendo defendido a dissertação em 20 de fevereiro de 2013.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

FREITAS, Leillimar dos Reis, M.Sc., Universidade Federal de Viçosa, fevereiro de 2013. **Comparação das funções de ligação logit e probit em regressão binária considerando diferentes tamanhos amostrais**. Orientador: Sebastião Martins Filho. Coorientadores: Fabyano Fonseca e Silva e José Ivo Ribeiro Júnior.

Considerou-se um estudo de regressão binária por meio as funções de ligação logit e probit visando verificar a robustez das funções de ligação diante da variação do tamanho da amostra. Estas funções de ligação utilizam, respectivamente, as distribuições acumuladas logística e normal, sendo a principal diferença entre elas os valores de probabilidades nos extremos da variável independente. Dentro desse contexto, foram realizadas simulações com 500 repetições utilizando amostras de 10 diferentes tamanhos, desde 10 a 91, com uma diferença entre as sucessivas amostras de 9 unidades. As medidas de desempenho percentual de convergência, erro quadrático médio da probabilidade geral, erro quadrático médio da probabilidade específica, teste Wald para os coeficientes, foram utilizadas para estabelecer uma recomendação para o uso das duas diferentes funções de ligação quando os dados foram gerados com o uso do logit e probit e analisados por ambas as funções de ligação em cada tamanho de amostra. Concluiu-se que o objetivo desse trabalho foi atingido ao estabelecer uma recomendação para o uso da função de ligação logit para tamanhos inferiores a 20 devido a maior taxa de convergência, ou seja, foi verificado com a utilização da função de ligação logit que há um maior número de amostras em que foi possível estimar os parâmetros da regressão binaria. Para maiores tamanhos de amostras, utilizando as demais medidas de desempenho, tanto o logit como o probit mostraram-se semelhantes, pois não foram encontradas diferenças significativas entre esses dois tipos de funções.

ABSTRACT

FREITAS, Leillimar dos Reis, M.Sc., Universidade Federal de Viçosa, february of 2013. **Comparison of logit and probit link functions in regression binary considering different sample sizes**. Adviser: Sebastião Martins Filho. Co-Advisers: Fabyano Fonseca e Silva and José Ivo Ribeiro Júnior.

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. These link functions apply, respectively, the cumulative distributions logistics and normal, and the probabilistic main difference values of independent variable extremes. Then, simulations were performed 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 concluded that the work aim was achieved by establishing a recommendation for the logit link function use for sizes below 20 due to higher convergence rate, ie, it was verified with logit link function that there is a greater number of samples which was possible to estimate the binary regression parameters. For larger sample sizes, using other performance measures, both, the logit and probit, were similar, as there were no significant differences between these two different functions.

INTRODUÇÃO

Nos modelos lineares de regressão, a variável dependente é expressa como uma função linear dos coeficientes de regressão. Há, no entanto, outras classes de modelos em que é possível escrevê-los mediante uma transformação nas variáveis.

Quando a variável dependente é do tipo qualitativa dicotômica, há a necessidade de abordar técnicas de regressão binária para o tratamento dos dados, uma vez que os modelos lineares não terão um bom ajuste. Além disso, um dos principais objetivos da regressão binária é estimar a probabilidade de ocorrência de determinado evento, ou seja, os resultados da variável dependente permite a interpretação em termos probabilísticos.

Para BENDER FILHO et al. (2010), uma maneira adequada de utilizar modelo baseados em escolhas qualitativas é pelas probabilidades, desse modo existem funções de ligações específicas como a logit e probit que com a utilização de funções de distribuições podem realizar o cálculo, essas funções possuem variável dependente binária. Mas quando essa variável assume mais que duas categorias, é importante utilizar outros métodos como o logit multimomial. De acordo com Barros (2008) a escolha da função de ligação logit assim como a probit é determinada por simples conveniência matemática e computacional

De acordo com a abordagem realizada por Cordeiro e Demétrio (2007), a função de ligação logit assim como a probit têm em comum o fato de a variável dependente ser uma variável qualitativa com dois possíveis valores; assim, as funções de ligação logit e probit são dadas respectivamente pelos inversos das distribuições acumuladas logística e normal. Devido à diferença nas formas das curvas representativas destas distribuições, é importante avaliar situações nas quais uma ou outra descrevem com precisão a probabilidade de interesse.

O presente trabalho teve como principal objetivo estabelecer recomendação quanto ao uso das funções de ligação logit e probit em função do tamanho da amostra. Utilizando 10 diferentes tamanhos amostrais e diversas medidas de desempenho, foi possível verificar diferenças entre as regressões que foram relevantes, de forma a estabelecer recomendações para o uso das funções logit e probit.

CAPITULO 1 – Regressão Binária

Este estudo aborda métodos de regressão binária simples, cujo principal objetivo é a realização do cálculo da probabilidade de se ter determinada característica. Este tipo de regressão possui a vantagem de ser mais flexível com relação a outros tipos. Assim, dentro desse contexto, foram utilizadas as funções de ligação logit e probit nas quais possuem como variável dependente binária. O logit e o probit utilizam de funções de distribuições especificas para a realização do cálculo da probabilidade, que são respectivamente, a logística e a normal. Os parâmetros das duas funções de ligação são estimados de forma iterativa pelo método da máxima verossimilhança, pois são transformações das distribuições acumuladas. Então, o principal objetivo deste capítulo é justificar o uso da regressão binária, e diferentes medidas de desempenho, com a abordagem teórica dos conceitos de regressão binária, métodos de estimação utilizados a fim de introduzir a aplicação prática.

1. Caracterização das variáveis

Quando se realiza algum estudo muitas vezes, as variáveis explicativas possuem natureza binária (presença ou ausência, aprovação ou reprovação, positivo ou negativo entre outras). Para Corrar et al. (2009), a variável dependente (Y), poderá assumir somente um de dois possíveis valores, chamados por conveniência de 0 ou 1; dessa forma, é possível calcular

$$P\left(Y=1 \mid X=x_i\right)=p_i \qquad \textbf{[1]}$$

e

$$P\left(Y=0 \mid X=x_i\right)=1-p_i \qquad \textbf{[2]}$$

como sendo a probabilidade de sucesso e fracasso, respectivamente, correspondente a cada nível x_i da covariável. Desse modo, o principal objetivo da análise estatística de regressão binária é investigar a relação entre a probabilidade de resposta e as variáveis explicativas.

Segundo Hair et al. (2009), a natureza da variável dependente binária (0 ou 1), viola os pressupostos de regressão linear; por exemplo, ausência de normalidade dos resíduos e a variância de uma variável dicotômica não é constante (heterocedasticidade). Assim, há uma família de modelos para dados categóricos como refere McCullagh e Nelder (1989), mais conhecidos como modelos lineares generalizados. O modelo mais conhecido é o modelo logístico, baseado na transformação logística da proporção; há ainda o probit que é uma alternativa quando a variável dependente também se apresenta de forma dicotômica.

2. Modelo linear generalizado

Para Casella e Berger (2010), a definição de um modelo linear generalizado é descrita por uma relação entre a média de uma variável resposta e uma variável dependente, Resende e Biele (2002) complementa que esses modelos, possuem como ideia principal ampliar as opções para a variável resposta, assim permitir flexibilidade para a relação entre a média da variável resposta e o preditor linear, ou seja, descreve uma relação entre E(Y) e X.

De acordo com Resende e Biele (2002), a técnica permite a generalização dos modelos lineares clássicos de variáveis contínuas, assim a estrutura para estimação dos modelos lineares normais pode ser estendida para modelos não lineares.

Segundo a abordagem realizada por Cordeiro e Demétrio (2007), as variáveis dependentes Y são estabelecidas assim que as observações a serem feitas são definidas, podendo ser contínuas ou discretas, com o ajuste de diferentes distribuições, com médias μ_i, isto é, $E(Y)= \mu_i$, $i=1,...,n$.

Cordeiro e Demétrio (2007) complementa que o modelo clássico de regressão é definido da seguinte forma:

$$\mathbf{Y}=\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \qquad [3]$$

em que Y é o vetor da variável dependente, $\boldsymbol{\mu}=E(\mathbf{Y})=\mathbf{X}\boldsymbol{\beta}$, componente sistemático, $\mathbf{X}$ é a matriz das variáveis independentes do modelo, $\boldsymbol{\beta}$ é o vetor de parâmetros, $\boldsymbol{\varepsilon}$ o componente aleatório com distribuição $\varepsilon_i \sim N(0, \sigma^2\mathbf{I})$. Assim, $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, $\mathbf{I}$ corresponde a matriz identidade e o vetor de médias $\boldsymbol{\mu}$ que define o componente aleatório, é igual ao preditor linear do componente sistemático.

Cordeiro e Demétrio (2007) acrescenta que existem casos em que não há a satisfação dessa estrutura entre o componente sistemático e o erro aleatório e não há motivos para restrição dessa estrutura, nem pela distribuição normal para o componente aleatório assim como a suposição de homogeneidade de variâncias. Ao longo dos anos outros modelos foram surgindo, desse modo um modelo linear generalizado é definido por uma distribuição para a variável dependente, um conjunto de variáveis independentes, cuja estrutura é linear e uma função de ligação entre a média da variável dependente e a estrutura linear.

De acordo com Cordeiro e Demétrio (2007), os modelos lineares generalizados podem ser utilizados quando existe uma única variável dependente Y associado a um conjunto de variáveis independentes com uma amostra de n observações. Agresti (1990) acrescenta que os modelos lineares generalizados são compostos de três componentes principais: um componente aleatório, que identifica a distribuição de probabilidades da variável resposta; um componente sistemático (modelo), que especifica a função linear de variáveis explicativas que são usadas como preditor; e, por fim, uma função de ligação, que descreve uma relação funcional entre o componente sistemático e o valor esperado do componente aleatório, resumindo, estabelece uma ligação entre os dois componentes. Assim os três componentes estão definidos da seguinte maneira:

i) O componente aleatório é representado por um conjunto de variáveis aleatórias dependentes com a mesma distribuição com médias $\mu_1,...,\mu_n$, isto é,

$$E(Y_i)=\mu_i \qquad i=1,...,n, \qquad \textbf{[4]}$$

em que

$$\begin{cases} Y_i = \{Y_1, Y_2, ..., Y_n\} \\ \mu_i = \{\mu_1, \mu_2, ..., \mu_n\} \end{cases}$$

ii) O componente sistemático, ou variáveis independentes ou explicativas (X) do modelo linear generalizado entram na forma de uma estrutura linear, é estabelecido durante o planejamento, essas variáveis entram na forma de uma soma linear, ou seja,

$$\boldsymbol{\eta}=\boldsymbol{X\beta} \qquad [5]$$

em que $\boldsymbol{X}=(X_1,...,X_n)^T$ é a matriz do modelo consistindo dos valores das variáveis independentes para as n observações, $\boldsymbol{\beta}=(\beta_1,...,\beta_p)^T$ é o vetor de parâmetros e o preditor linear é dado pelo vetor $\boldsymbol{\eta}=(\eta_1,...,\eta_n)^T$.

iii) A Função de ligação é o terceiro componente do modelo linear generalizado, é a função de ligação que depende do tipo de resposta ou da aplicação. Demétrio (2002) informa que uma função de ligação deve satisfazer a condição de transformar o intervalo (0,1) em valores reais. De um modo geral, relaciona a média ao preditor linear, ou seja, estabelece uma relação linear direta, isto é,

$$\eta_i = g(\mu_i) \qquad i=1,..., n.$$

em que $g(\mu_i)$ é uma função de ligação. As funções de ligação para o logit e probit são respectivamente dadas por:

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) = g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

e

$$\eta = \Phi^{-1}(\mu) = g(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i$$

em que $\mu = E(Y_i)$.

A definição abordada por Cordeiro e Demétrio (2007) sobre modelos lineares generalizados informa que não há uma aditividade entre o erro aleatório ε e a média μ como no modelo clássico de regressão, produzindo o componente aleatório. Desse modo, no modelo linear generalizado define-se uma distribuição para a variável dependente que representa os dados e não uma distribuição para o erro.

3. Regressão Binária

Segundo Cordeiro e Demétrio (2007), dentre os métodos estatísticos para a análise de dados que são casos especiais de modelos lineares generalizados há o logit e o probit. Stock e Watson (2004) acrescenta que a função de ligação logit é semelhante à probit exceto pela substituição da função de distribuição acumulada utilizada para a realização do cálculo da probabilidade, ou seja, enquanto que a função de ligação logit utiliza da distribuição logística, a função de ligação probit utiliza da distribuição normal, isto é,

$$p_i = F(\beta_0 + \beta_1 x_i) \qquad \textbf{[6]}$$

De acordo com Gujarati (2005), efetua-se uma transformação na variável dependente para o uso da função de ligação logit, cuja primeira etapa consiste em convertê-la em uma chance, isto é,

$$p_i = P\left(Y = 1 \mid X = x_i\right) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} = \frac{1}{1+e^{-\eta_i}}. \qquad \textbf{[7]}$$

A equação 7 é conhecida como função logística acumulada.

Como se pode observar η_i varia de -∞ a +∞, $0 \le p_i \le 1$ e ainda que p_i não se relaciona linearmente com η_i; como p_i é não linear não somente em X, mas também nos β's, portanto o método dos mínimos quadrados ordinários não pode ser utilizado.

Se a probabilidade de possuir determinada característica é dada por $P\left(Y = 1 \mid X = x_i\right) = p_i$, a probabilidade de não possuir será dada pela expressão 8.

$$1 - p_i = 1 - \frac{1}{1+e^{-\eta_i}} = \frac{1+e^{-\eta_i}-1}{1+e^{-\eta_i}} = \frac{e^{-\eta_i}}{1+e^{-\eta_i}}\frac{e^{\eta_i}}{e^{\eta_i}} = \frac{1}{1+e^{\eta_i}}. \qquad \textbf{[8]}$$

Portanto,

$$\frac{p_i}{1-p_i} = \frac{\dfrac{1}{1+e^{-\eta_i}}}{\dfrac{1}{1+e^{\eta_i}}} = \frac{1+e^{\eta_i}}{1+e^{-\eta_i}} = \frac{1+e^{\eta_i}}{1+\dfrac{1}{e^{\eta_i}}} = \frac{1+e^{\eta_i}}{\dfrac{e^{\eta_i}+1}{e^{\eta_i}}} = \frac{1+e^{\eta_i}}{1}\frac{e^{\eta_i}}{1+e^{\eta_i}} = e^{\eta_i} \qquad \textbf{[9]}$$

Desse modo, se p_i =0,2 tem-se chances de 1 para 4.

A função de ligação probit utiliza da função de distribuição acumulada normal, ou seja,

$$P\left(Y = 1 \mid X = x_i\right) = \Phi\left(\beta_0 + \beta_1 x_i\right) \qquad \textbf{[10]}$$

em que Φ é a função de distribuição acumulada normal padrão. Se β_1 for positivo, um aumento em x, aumentará a probabilidade de Y=1, caso contrário, um aumento em x diminuirá a probabilidade de Y=1.

De acordo com Stock e Watson (2004), a expressão $\beta_0+\beta_1 x_i$ no probit, desempenha o papel de "z", na tabela de distribuição acumulada normal padrão.

Uma das mais importantes funções de ligação é baseada na transformação logit e probit para proporção; para se evitar que os valores das probabilidades se situem fora do intervalo [0,1] é efetuada uma transformação onde as funções de ligação logit e probit são dadas respectivamente por:

$$g(p_i) = F^{-1}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_i \qquad \textbf{[11]}$$

e

$$g(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i\,, \qquad \textbf{[12]}$$

ou seja, as funções de ligação são dadas pelas inversas das distribuições acumuladas associadas.

A principal diferença entre estas duas distribuições (logística e normal) está nas probabilidades referentes aos valores extremos da covariável, ou seja, no peso das suas caudas sendo que as principais semelhanças estão nas formas das curvas (campanular), simetria e que $f(x)$ tende a zero quando x tende a $\pm\infty$ (assintótica com relação ao eixo x), como pode ser observado na Figura 1.

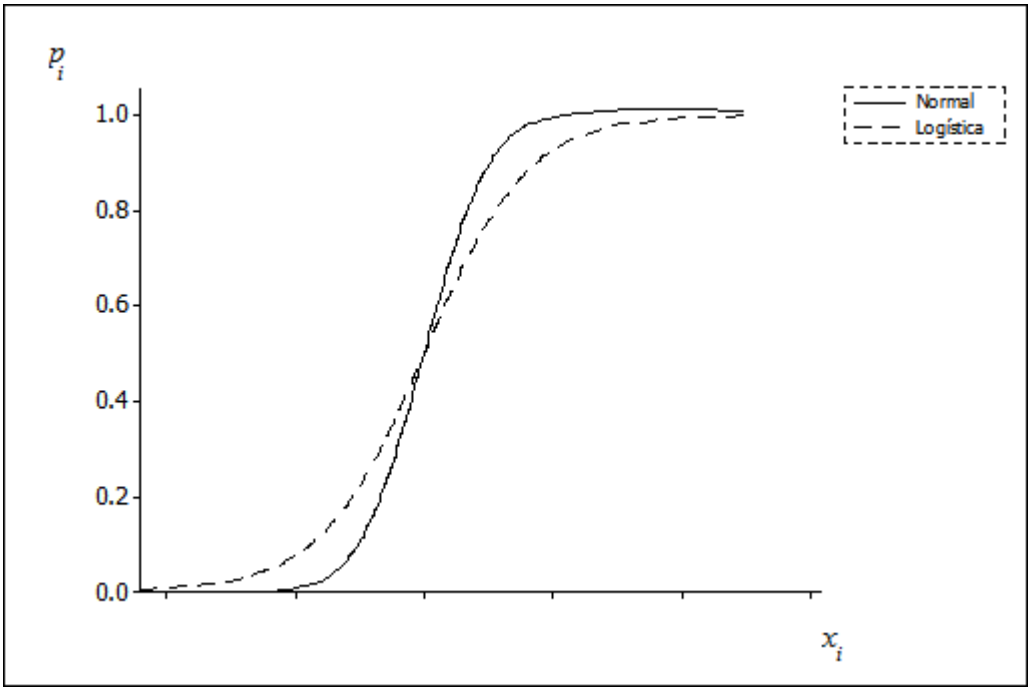

Figura 1– Comparação gráfica das distribuições acumulada logística e normal.

De acordo com Corrar et al. (2009), quando tem-se um modelo linear, uma das alternativas para se estimar os parâmetros é o metodo dos mínimos quadrados, mas no caso do logit e probit, deve-se recorrer a outro método, conhecido como método da máxima verossimilhança.

Segundo Cramer (2003), os primeiros trabalhos publicados sobre logit foram feitos no final das décadas de 1950 e 1960 em estatística e epidemiologia; na estatística havia uma vantagem analítica na transformação do logit em lidar com saídas binárias, uma vez, que todos os cálculos eram realizados a mão. Na epidemiologia o estudo do logit se deu ainda mais cedo (1950), uma vez que estava diretamente ligada à razão de chances de probabilidades. Corrar et al. (2009) acrescentam que essa técnica foi desenvolvida para tentar realizar predições ou tentar explicar a ocorrência de determinados fenômenos quando a variável dependente é de natureza binária.

Corrar et al. (2009) informam um dos motivos que as funções de ligação vêm sendo largamente utilizadas, para realizar predições quando a variável dependente é dicotômica, é devido ao pequeno número de restrições que são elas: incluir todas as variáveis para que se obtenha maior estabilidade; valor esperado do erro deve ser zero; inexistência de autocorrelação entre os erros; inexistência de correlação entre os erros e as variáveis independentes e; ausência de multicolinearidade perfeita entre as variáveis independentes.

Os últimos autores citados acrescentam que existe um problema quando não se tem variáveis independentes normais no caso linear, mas como a variável dependente é do tipo dicotômica (com distribuição de Bernoulli) e no caso das funções de ligação logit e probit não há essa restrição. Quanto ao número de observações necessárias para se realizar inferências de boa qualidade, não há na literatura, de acordo com Corrar et al. (2009), um consenso. Assim, os autores informam que quando se trabalha com o logit, devem-se obter amostras maiores que no caso linear, mas essas funções de ligação possuem a vantagem de acolher mais facilmente variáveis dependentes binárias.

Para explicar o sucesso da regressão binária, Corrar et al. (2009), atribuem os seguintes fatores: acolhe com maior facilidade as variáveis categóricas; uma das alternativas é a análise discriminante principalmente no que se refere a problemas com homogeneidade de variâncias porém, essa alternativa possui fortes pressuposições como ausência de pontos discrepantes, normalidade e homogeneidade das variâncias e covariancias; porém é mais adequada a solução de problemas que envolve a estimação de probabilidades. Pereira et al. (2007) acrescenta que o modelo logit é mais robusto que a análise discriminante, uma vez que se aplica a

distribuições não normais. Se comparado com o probit, o logit tem representação e tratamento matemático mais simples, justificando a sua maior utilização.

Para Cramer (2003), a criação do probit é creditada a Gaddum e Bliss, mas Fechner, um estudioso alemão, foi o primeiro a transformar diferenças observadas equivalentes ao desvio normal. O termo probit foi introduzido por Gaddum e Bliss, que significa unidade de probabilidade, pois em seus escritos, quando iniciou o bio-ensaio ambos os autores aderiram firmemente ao modelo clássico, onde o estímulo era determinístico e respostas aleatórias, por causa da variabilidade dos níveis de tolerância individual, mas após um ano, essa teoria foi abandonada.

De acordo com Cramer (2003), sem a teoria do bioensaio, o probit foi rapidamente difundido para qualquer relação que descrevesse um resultado binário discreto a uma ou mais variáveis resposta. Na economia, por exemplo, o probit foi utilizado pela primeira vez na década de 1950.

Cramer (2003) complementa que, ao longo dos anos, o número de trabalhos publicados referentes ao logit teve rápido crescimento se comparado ao probit; o que se deve principalmente à facilidade de se realizar cálculos sem o uso computacional, uma vez que até aproximadamente 1980, a questão computacional era uma questão importante no que se refere ao uso de cálculos para a realização da estimação.

O método de estimação utilizado pelo logit e probit, segundo Stock e Watson (2004), é o metodo da máxima verossimilhança, pois produzem estimadores eficientes (variância mínima), consistentes e normalmente distribuídos para grandes amostras, de forma que diversas estatísticas, como o intervalo de confiança, podem ser obtidas de forma usual.

4. Método de estimação da regressão binária

De acordo com Hair et al. (2009), a regressão linear utiliza dos métodos dos mínimos quadrados ordinários para realizar a estimação de seus coeficientes, esse método consiste em minimizar a soma de quadrados das diferenças entre os valores observados e os previstos. Na regressão não linear o método da máxima verossimilhança é utilizado de forma iterativa para que sejam encontradas as estimativas mais prováveis dos parâmetros. Ao invés de minimizar os desvios

quadrados, a regressão não linear maximiza a probabilidade de que um evento ocorra.

Casella e Berger (2010) complementam que quando se usa regressão linear, a técnica de mínimos quadrados é uma opção para o cálculo dos estimadores; nos modelos não lineares não há uma conexão direta entre a variável dependente (Y_i) e o componente sistemático ($\beta_0+\beta_1 x_i$), assim o metodo dos mínimos quadrados não é mais uma opção, sendo a estimação realizada por meio do método da máxima verossimilhança.

Lemonte (2006) acrescenta que, muito frequentemente, as observações retiradas de uma população com uma função de densidade de probabilidade $F(y, \beta)$ são mutuamente independentes para todas as distribuições, então a função de verossimilhança $L(\beta, y)$ do vetor de parâmetros **β** pode ser escrita como um produtório,

$$L(\beta \mid \boldsymbol{y}) = L(\beta_0, \beta_1 \mid y_1, ..., y_n) = \prod_{i=1}^{n} F(y_i \mid \beta_0, \beta_1) = \prod_{i=1}^{n} F(y_i, \beta_i), \qquad \textbf{[13]}$$

ou seja, a função de densidade de probabilidade conjunta $F(y_i,\beta_i)$ é o produto das densidades de cada uma das observações. A interpretação da função de densidade de probabilidade conjunta pode ser descrita como uma função em que o vetor de parâmetros se torna variáveis.

Segundo Casella e Berger (2010) o método da máxima verossimilhança é definido como sendo os valores dos parâmetros que geram, com maior frequência, a amostra observada. Para a realização do procedimento, deve-se maximizar a função de verossimilhança com relação à $\hat{\beta}$, assim iguala-se a zero as derivadas parciais da função de verossimilhança e determinar $\hat{\beta}$ que solucione o conjunto de equações. Então, para facilitar o manuseio da equação trabalha-se com o logaritmo natural da função de verossimilhança (lnL), pois maximizar o logaritmo natural de uma função é, em geral, mais simples e produz os mesmos resultados da maximização da função original. Logo, deve-se resolver o sistema $\boldsymbol{U}_j = \partial \ln L / \partial \beta_j = 0$ para obter a função escore.

Segundo a abordagem realizada por Demétrio (2002), as equações U_j=0, j=1,2,... não são lineares e devem ser resolvidas por processos iterativos do tipo Newton-Raphson. O método iterativo de Newton-Raphson para a solução de uma

dada equação $F(x)=0$ é baseado na aproximação de Taylor para a função $F(x)$ nas vizinhanças do ponto x_0.

Para obter a solução do sistema $U_j = \partial \ln L / \partial \beta_j = 0$, Demétrio (2002) utiliza da versão multivariada do método de Newton-Raphson, então

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \left(\boldsymbol{I}_0^{-1}\right)^{(k)} \boldsymbol{U}^{(k)} \qquad \textbf{[14]}$$

sendo $\boldsymbol{\beta}_k$ e $\boldsymbol{\beta}_{k+1}$ os vetores de parâmetros estimados nos passos k e $(k+1)$, o vetor escore (vetor de derivadas parciais de $f(x)$), com elementos $\partial l/\partial \beta j$, avaliado no passo k e $\left(\boldsymbol{I}_0^{-1}\right)^{(k)}$ a inversa da negativa da matriz de derivadas parciais de segunda ordem de $F(x)$, com elementos $\partial^2 l/\partial \beta_j \partial \beta_i$, avaliada no passo k.

Demétrio (2002) acrescenta que, se as derivadas de segunda ordem são obtidas facilmente, o método de Newton-Raphson é útil. Mas, isso nem sempre ocorre, assim, no caso dos modelos lineares generalizados utiliza-se o método escore de Fisher, que envolve a substituição da matriz de derivadas parciais de segunda ordem pela matriz de valores esperados das derivadas parciais, ou seja, a substituição da matriz de informação observada, $\boldsymbol{I}_0$, pela matriz de informação esperada de Fisher, $\mathfrak{I}$. Logo,

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + \left(\mathfrak{I}^{-1}\right)^{(k)} \boldsymbol{U}^{(k)} \qquad \textbf{[15]}$$

cujos elementos de $\mathfrak{I}$ é dado por

$$\mathfrak{I}_{jk} = E\left[\frac{-\partial \ell}{\partial \beta_i \partial_j}\right] = E\left[\frac{\partial \ell}{\partial \beta_i}\frac{\partial \ell}{\partial_j}\right] \qquad \textbf{[16]}$$

que é a matriz de covariâncias dos $\boldsymbol{U}'_j$.

Os estimadores de máxima verossimilhança possuem algumas propriedades ótimas, como não tendenciosidade ($E\left(\widehat{\beta}-\beta\right)=0$), consistência ($\lim_{n\to\infty} E\left(\widehat{\beta}-\beta\right)=0$) e eficiência ($\lim_{n\to\infty} var\left(\widehat{\beta}\right)=0$).

Para Casella e Berger (2010) quando se realiza uma amostragem a partir de uma população descrita por uma função de probabilidade (f.p.) ou por uma função

densidade de probabilidade (f.d.p.), o conhecimento do estimador, dado por $\hat{\theta}$, gera o conhecimento de toda a população, assim é necessário encontrar um bom estimador; para estimadores pontuais qualquer estatística é um estimador. Dentro da classe de estimação pontual um dos métodos existentes é o dos mínimos quadrados ordinários e da máxima verossimilhança

Os métodos de estimação possuem algumas propriedades no que se refere aos estimadores. Segundo Bolfarine e Sandoval (2000), essas propriedades são eficiência (que são obtidos apenas pela família exponencial de distribuição); um estimador para ser ótimo, de acordo com o critério do menor erro quadrático médio, deve ser função de uma estatística suficiente (são aquelas que resumem os dados sem perder nenhuma informação, elas são tão informativas quanto à amostra toda).

Para uma melhor escolha dos estimadores, Magalhães e Lima (2008), informam que é importante eles possuirem as propriedades de ser não viciado (viesado) e consistente. Um estimador $\hat{\theta}$, por exemplo, é dito não viciado se o valor esperado é igual ao observado, ou seja, se $E\left(\hat{\theta}\right)=\theta$. E, um estimador é dito consistente se, na medida em que o tamanho da amostra aumenta, o valor esperado do estimador converge para o parâmetro de interesse e sua variância converge para zero, ou seja, se $\lim_{n\to\infty} E\left(\hat{\theta}\right)=\theta$ e $\lim_{n\to\infty} Var\left(\hat{\theta}\right)=0$. Assim, pode-se perceber que a consistência depende do tamanho da amostra, o vício o deve valer para qualquer tamanho de *n*.

Magalhães e Lima (2008) complementam que quando dois estimadores forem consistentes e não viciados para um parâmetro, pode-se utilizar o conceito de eficiência. Considerando dois estimadores, $\hat{\theta}_1$ e $\hat{\theta}_2$, não viciados para o parâmetro θ, pode-se dizer que $\hat{\theta}_1$ é mais eficiente do que $\hat{\theta}_2$ se $\text{var}\left(\hat{\theta}_1\right) < \text{var}\left(\hat{\theta}_2\right)$.

5. Teste de Wald

De acordo com Demétrio (2002), existem três estatísticas para testar os parâmetros da regressão binária que são: teste da razão de verossimilhança, teste de Wald e teste escore. O autor complementa que essas estatísticas são assintoticamente equivalentes, sendo que o teste da razão de verossimilhança (TRV) é definido como

o mais poderoso, ou seja, há um maior aumento da probabilidade de rejeição da hipótese nula dado que ela é falsa do teste TRV com relação ao teste Wald; porém a estatística do teste TRV utilizada é qui-quadrado, portanto requer um tamanho de amostra maior.

O teste Wald (ou teste de Wald), para Hair et al. (2009) é parecido com os valores *F* ou *t* para o teste de significância dos coeficientes na regressão linear. Quando os coeficientes são significantes sua interpretação é que as variáveis podem ser utilizadas para identificar às relações que afetam as probabilidades previstas. A mesma interpretação pode ser realizada para a constante. Desse modo, a hipótese nula a ser testada é que

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

Segundo Corrar et al. (2009), a finalidade deste teste é verificar o grau de significância para cada coeficiente da equação, ou seja, se cada parâmetro é significativamente diferente de zero, mais especificamente, verifica a hipótese de que um determinado coeficiente é igual à zero. Essa estatística pode ser calculada do seguinte modo:

$$W_{calc} = \frac{\hat{\beta}}{S\left(\hat{\beta}\right)} \qquad \textbf{[17]}$$

em que β é definido como sendo a estimativa do coeficiente de uma variável independente incluída no modelo; e $S\left(\hat{\beta}\right)$ é o erro padrão que é definido da seguinte forma:

$$S^2\left(\hat{\beta}\right) = -\left[E\left(\Im\left(\beta\right)\right)\right]^{-1} \qquad \textbf{[18]}$$

em que $\Im$ é a matriz de informação de Fisher.

O p-valor é definido como $P(|Z|>|W_{calc}|)$, sendo que Z corresponde a variável aleatória da distribuição normal padrão.

6. Erro quadrático médio

Segundo Lira (2008), o erro quadrático médio avalia a qualidade do estimador ($\hat{\theta}$); ele evidencia duas componentes de variabilidade dos dados, a variância do estimador (precisão) e o vício (acurácia).

O erro quadrático médio de um estimador é definido por Bolfarine e Sandoval (2000) da seguinte maneira:

$$EQM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta}-\theta\right)^2\right]. \qquad \textbf{[19]}$$

Resolvendo a equação anterior:

$$EQM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta}-E\left(\hat{\theta}\right)+E\left(\hat{\theta}\right)-\theta\right)\right]^2$$

$$EQM\left[\hat{\theta}\right] = E\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]^2 + 2E\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]\left[E\left(\hat{\theta}\right)-\theta\right]+\left[E\left(\hat{\theta}\right)-\theta\right]^2, \text{ mas}$$

$$E\left(\hat{\theta}-\theta\right)E\left(\hat{\theta}-E\left(\hat{\theta}\right)\right) = E\left(\hat{\theta}\right)-E\left[E\left(\hat{\theta}\right)\right] = E\left(\hat{\theta}\right)-E\left(\hat{\theta}\right) = 0, \text{ portanto}$$

$$EQM\left[\hat{\theta}\right] = E\left[\hat{\theta}-E\left(\hat{\theta}\right)\right]^2 + \left[E\left(\hat{\theta}\right)-\theta\right]^2, \qquad \textbf{[20]}$$

ou seja, $EQM\left[\hat{\theta}\right] = Var\left(\hat{\theta}\right)+\left[E\left(\hat{\theta}\right)-\theta\right]^2$, em que $E\left(\hat{\theta}\right)-\theta = B$, sendo B o vício do estimador. O EQM muitas vezes se mostra melhor do que a variância quando o vício não é desprezível, pois é dado pela soma dessas duas estatísticas.

De acordo com Lira (2008), a raiz quadrada da variância é chamada de erro padrão, isto é, $EP\left(\hat{\theta}\right) = \sqrt{Var\left(\hat{\theta}\right)}$, quanto menor o erro padrão, maior a precisão das estimativas.

Uma medida muito utilizada na estatística é o coeficiente de variação, isto é, precisão relativa. Esta precisão é dada pelo inverso do coeficiente de variação

(*CV*), ou seja, quanto maior *CV*, menor é a precisão, assim o coeficiente de variação é calculado por:

$$CV\left(\hat{\theta}\right) = \frac{\sqrt{Var\left(\hat{\theta}\right)}}{E\left(\hat{\theta}\right)} = \frac{EP\left(\hat{\theta}\right)}{E\left(\hat{\theta}\right)}$$

Lira (2008) complementa que um estimador é dito não viciado se $E\left[\hat{\theta}\right] = \theta$. Desta forma, o erro quadrático médio é a soma da variância e do quadrado do vício (viés), cujo $\hat{\theta}$ é definido como as estimativas dos parâmetros da equação ou a estimativa da probabilidade calculada, isto é, $\hat{\theta} = \hat{\beta}_i$ ou $\hat{\theta} = \hat{p}_i$. Então, o *EQM* possui algumas propriedades ótimas dos estimadores como a não tendenciosidade ($E\left[\hat{\theta}\right] = \theta$), consistência ($\lim_{n\to\infty} E\left[\hat{\theta}\right] = \theta$ e $\lim_{n\to\infty} var\left[\hat{\theta}\right] = \theta$) e eficiência.

CAPITULO 2 – Funções de ligação logit e probit na regressão binária via simulação de dados

RESUMO: Neste estudo foi considerada a regressão binária por meio as funções de ligação logit e probit, visando verificar a robustez das funções de ligação diante da variação do tamanho da amostra. Assim, foram realizadas simulações com 500 repetições utilizando amostras de 10 diferentes tamanhos, desde 10 a 91, com uma diferença entre as sucessivas amostras de 9 unidades. As medidas de desempenho percentual de convergência, erro quadrático médio da probabilidade geral, erro quadrático médio da probabilidade específica, teste Wald para os coeficientes, foram utilizadas para estabelecer uma recomendação para o uso das duas diferentes funções de ligação quando os dados foram gerados com o uso do logit e probit e analisados por ambas as funções de ligação em cada tamanho de amostra. Concluiu-se que o objetivo desse trabalho foi atingido ao estabelecer uma recomendação para o uso da função de ligação logit para tamanhos inferiores a 20 devido a maior taxa de convergência. Para maiores tamanhos de amostras, utilizando as demais medidas de desempenho, tanto o logit como o probit mostraram-se semelhantes.

Palavras Chave: tamanho de amostra, variável binária, distribuições logística e normal.

1. Introdução

Muitos modelos são casos especiais de modelos lineares generalizados que são compostos de três componentes: um componente aleatório (identifica a distribuição de probabilidades da variável dependente); um componente sistemático (modelo – especifica a função linear de variáveis explicativas que são usadas como preditor); e por uma função de ligação (estabelece uma ligação entre os dois componentes).

Nos modelos lineares de regressão, a variável dependente é expressa como uma função linear dos coeficientes de regressão. Há, no entanto, outras classes de modelos em que é possível escrevê-los mediante uma transformação nas variáveis.

Quando a variável dependente é do tipo qualitativa dicotômica, há a necessidade de abordar técnicas de regressão binária para o tratamento dos dados, uma vez que os modelos lineares não terão um bom ajuste. Além disso, um dos principais objetivos da regressão binária é estimar a probabilidade de ocorrência de determinado evento, ou seja, os resultados da variável dependente permitiram a interpretação em termos de probabilísticos.

Para BENDER FILHO et al. (2010), uma maneira adequada de utilizar modelo baseados em escolhas qualitativas é pelas probabilidades, desse modo existem funções de ligações específicas como o logit e probit que com a utilização de funções de distribuições podem realizar o cálculo, essas funções possuem variável dependente binária. Mas quando essa variável assume mais que duas categorias, é importante utilizar outros métodos como o logit multimomial. De acordo com Barros (2008) a escolha da função de ligação logit assim como a probit é determinada por simples conveniência matemática e computacional

De acordo com a abordagem realizada por Cordeiro e Demétrio (2007), a função de ligação logit assim como a probit têm em comum o fato de a variável dependente ser uma variável qualitativa com dois possíveis valores; assim, as funções de ligação logit e probit são dadas respectivamente pelos inversos das distribuições acumuladas logística e normal. Devido à diferença nas formas das curvas representativas destas distribuições, é importante avaliar situações nas quais uma ou outra descrevem com precisão a probabilidade de interesse.

O presente trabalho teve como principal objetivo verificar o efeito do tamanho da amostra sobre a qualidade de ajuste e da robustez das funções de ligação logit e probit, quando a variável dependente dicotômica é originada de uma variável latente que assume distribuições de probabilidade logística e normal e; estabelecer recomendações para escolha das funções de ligação logit e probit ao ajuste da regressão de uma variável dependente dicotômica em função do tamanho da amostra.

Então, utilizando de 10 diferentes tamanhos amostrais e diversas medidas de desempenho, foi possível verificar diferenças entre as regressões que foram relevantes, de forma foi estabelecida recomendações para o logit e probit. Assim, espera-se que esse trabalho possa contribuir para a escolha dos tipos de função de ligação em função de diferentes tamanhos de amostras.

2. Material e Métodos

2.1. Simulação de dados

Para a realização da simulação, inicialmente foram definidos o tamanho da amostra, qual o tipo de equação foi 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 ($1 \leq x \leq 10$)	x	Tamanhos das amostras (n)
10	1; 2; 3;...; 10	10
20	1; 1;5; 2; ...; 10	19
30	1; 1;33; 1;67; 2; ...; 10	28
40	1; 1;25; 1;5; ...; 10	37
50	1;0; 1;2; 1;4; 1;6;...; 10	46
60	1; 1;167; 1;33;...; 10	55
70	1; 1;14; 1;28;...; 10	64
80	1; 1;125; 1;250;...; 10	73
90	1; 1;11; 1;22;...;10	82
100	1; 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, a 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:

$$\begin{cases} logit_i = g(p_i) = \beta_0 + \beta_1 x_i \\ probit_i = g(p_i) = \beta_0 + \beta_1 x_i \end{cases} \quad \textbf{[1]}$$

em que esta equação foi considerada como verdadeira 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 $1 \leq x \leq 10$.

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 $P(Y=1|X=x_i) = p_i$ (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 $Y=1|X=x_i$ de acordo expressões 2 e 3:

$$p_i = P(Y=1 \mid X=x_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}, \text{ para } 1 \leq x_i \leq 10 \quad \textbf{[2]}$$

$$p_i = P(Y=1 \mid X=x_i) = \Phi(\beta_0 + \beta_1 x_i), \text{ para } 1 \leq x_i \leq 10 \quad \textbf{[3]}$$

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. Portanto, tem-se:

$$Y|x_i \sim \text{Ber}(p_i), \text{ para } p_i = p_{Li} \text{ e } p_i = p_{Pi}. \quad \textbf{[4]}$$

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. Isto 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, 2012). 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, ou seja, foi realizado um total de 10.000 análises, isto é, 500 simulações x10 tamanhos de amostra x 2 funções de ligação.

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).

2.2. 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. Portanto, 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 ditos pela função de ligação na qual os dados tiveram origem, mas por outro lado, por meio do outro tipo de função de ligação (probit). O mesmo aconteceu quando a análise de regressão foi realizada por meio da função de ligação probit (Figura 2).

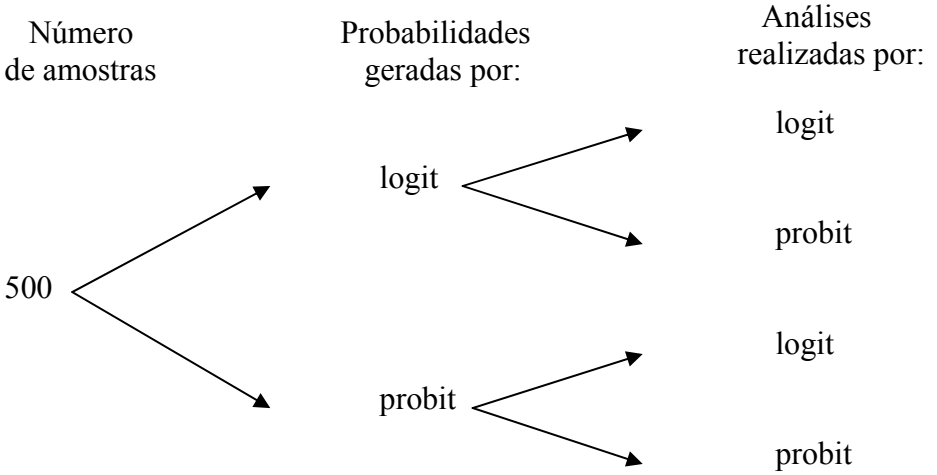

Figura 2 – Esquema das análises realizadas utilizando as funções de ligação logit e probit para cada tamanho de amostra

2.3. 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 algumas medidas de desempenho: percentual de convergência, erro quadrático médio da probabilidade geral estimada em relação à verdadeira, erro quadrático médio da probabilidade específica estimada em relação à verdadeira e teste de Wald dos parâmetros.

i) Percentual de convergência: é a 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;

ii) Erro quadrático médio da probabilidade geral estimada em relação à verdadeira: o cálculo dessa estatística foi obtido com a utilização de todos os diferentes valores de x ($1 \leq x \leq 10$), ou seja,

$$EQM\left[\hat{p}\right] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{500}\left(\hat{p}_{ij} - p_{ij}\right)^2}{500n} \qquad \textbf{[5]}$$

em que n=10,19,28,...,91, $\hat{\beta}$ é o valor assumido pela estatística, e β são os valores verdadeiros da constante (β_0=-5,5) e do coeficiente (β_1=1);

iii) Erro quadrático médio da probabilidade específica estimada em relação à verdadeira: seu cálculo foi obtido com a utilização dos níveis específicos de $1 \leq x \leq 10$, ou seja, x iguais a 1, 2, 3, 4, 5, 6, 7, 8, 9 e10

$$EQM\left[\hat{p}\right]_{x_1} = \frac{\sum_{j=1}^{500}\left(\hat{p}_{ij} - p_{ij}\right)^2}{500}, \qquad \textbf{[6]}$$

$$EQM\left[\hat{p}\right]_{x_2} = \frac{\sum_{j=1}^{500}\left(\hat{p}_{ij} - p_{ij}\right)^2}{500}, \qquad \textbf{[7]}$$

$$\vdots$$

$$EQM\left[\hat{p}\right]_{x_{10}} = \frac{\sum_{j=1}^{500}\left(\hat{p}_{ij} - p_{ij}\right)^2}{500}; \qquad \textbf{[8]}$$

iv) Teste de Wald 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); Então o teste verificou a significância das seguintes hipóteses:

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

v) Análise de regressão: depois de obtidos os 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 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,

$$md_{\text{logit}} = \lambda_0 + \lambda_1 n + \lambda_2 n^2 + \lambda_3 f + \lambda_4 nf + \varepsilon \quad \textbf{[9]}$$

e

$$md_{\text{probit}} = \gamma_0 + \gamma_1 n + \gamma_2 n^2 + \gamma_3 f + \gamma_4 nf + \varepsilon \quad \textbf{[10]}$$

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

3. Resultados e Discussão

3.1. Percentual de Convergência

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

c = -38,8778 + 7,17778* n, para $10 \leq n < 19$ **[11]**

c = 99,73, para $19 \leq n \leq 91$ **[12]**

ou seja, o tamanho da amostra influencia 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 (Figura 3).

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.

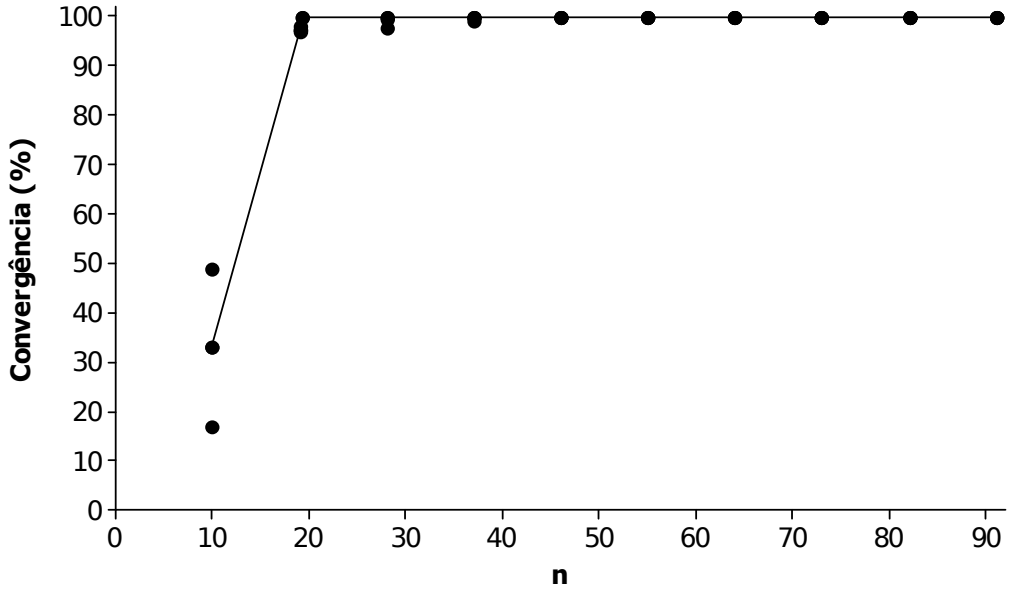

Figura 3 – Percentuais de convergências do logit e probit

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.

3.2. Erro quadrático médio da probabilidade geral

O erro quadrático médio diminuiu ($P<0,05$) 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 ($P>0,05$) entre as funções

logit e probit. Os parâmetros nas equações (Figura 4 a e b) foram significativos pelo teste t de Student (P<0,05).

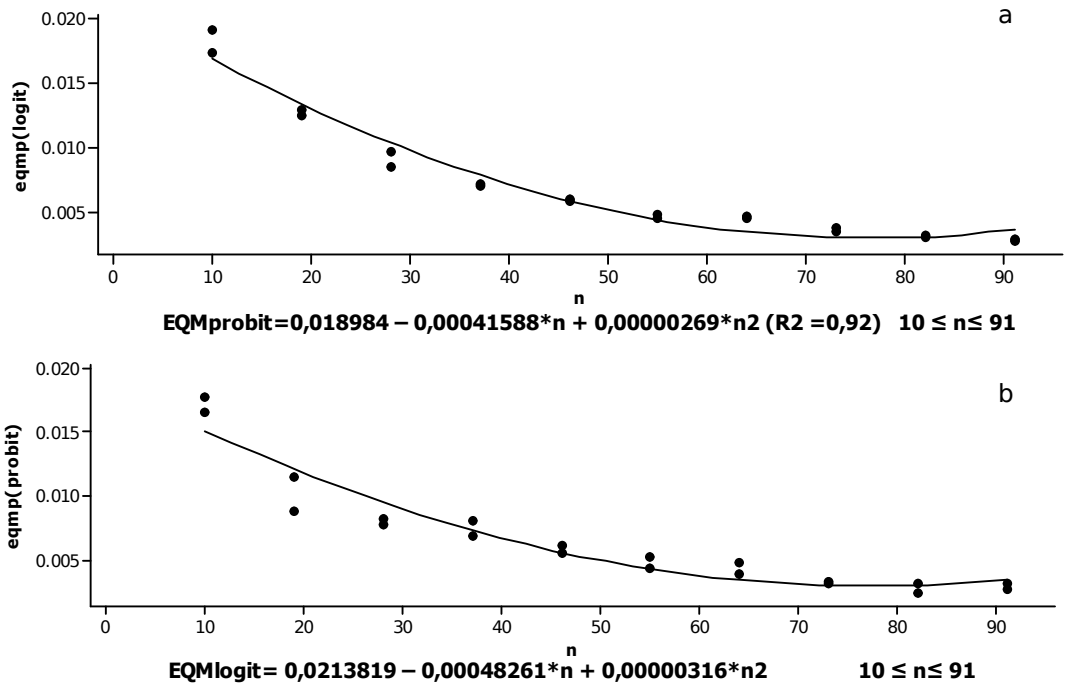

Figura 4 – Erro quadrático médio dos dados oriundos das funções de ligação logit e probit, em função do tamanho da amostra (*n*).

Segundo Miot (2011), o erro é inversamente proporcional ao tamanho da amostra, como pode ser observado na Figura 4, 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.

3.3. 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<0,05). Além disso, não houve diferença entre os dois tipos de funções de ligação empregadas (P>0,05).

Para os dados simulados a partir da função de ligação logit fixados diferentes níveis de x, as equações de regressão ajustadas estão apresentadas na Tabela 3 e as curvas na Figura 5.

Tabela 3 – Equação de regressão e grau de ajustamento para o erro quadrático médio 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	$E\hat{Q}M = 0{,}00420 - 0{,}000132^* n + 0{,}000001^* n^2$	0,64
x=2	$E\hat{Q}M = 0{,}00693 - 0{,}000216^* n + 0{,}000002^* n^2$	0,60
x=3	$E\hat{Q}M = 0{,}0196 - 0{,}000510^* n + 0{,}000004^* n^2$	0,57
x=4	$E\hat{Q}M = 0{,}0286 - 0{,}000590^* n + 0{,}000004^* n^2$	0,84
x=5	$E\hat{Q}M = 0{,}0412 - 0{,}000903^* n + 0{,}000006^* n^2$	0,60
x=6	$E\hat{Q}M = 0{,}0499 - 0{,}00109^* n + 0{,}000007^* n^2$	0,75
x=7	$E\hat{Q}M = 0{,}0318 - 0{,}000826^* n + 0{,}000006^* n^2$	0,62
x=8	$E\hat{Q}M = 0{,}0151 - 0{,}000386^* n + 0{,}000003^* n^2$	0,58
x=9	$E\hat{Q}M = 0{,}00432 - 0{,}000105^* n + 0{,}000001^* n^2$	0,76
x=10	$E\hat{Q}M = 0{,}00981 - 0{,}000342^* n + 0{,}000003^* n^2$	0,51

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

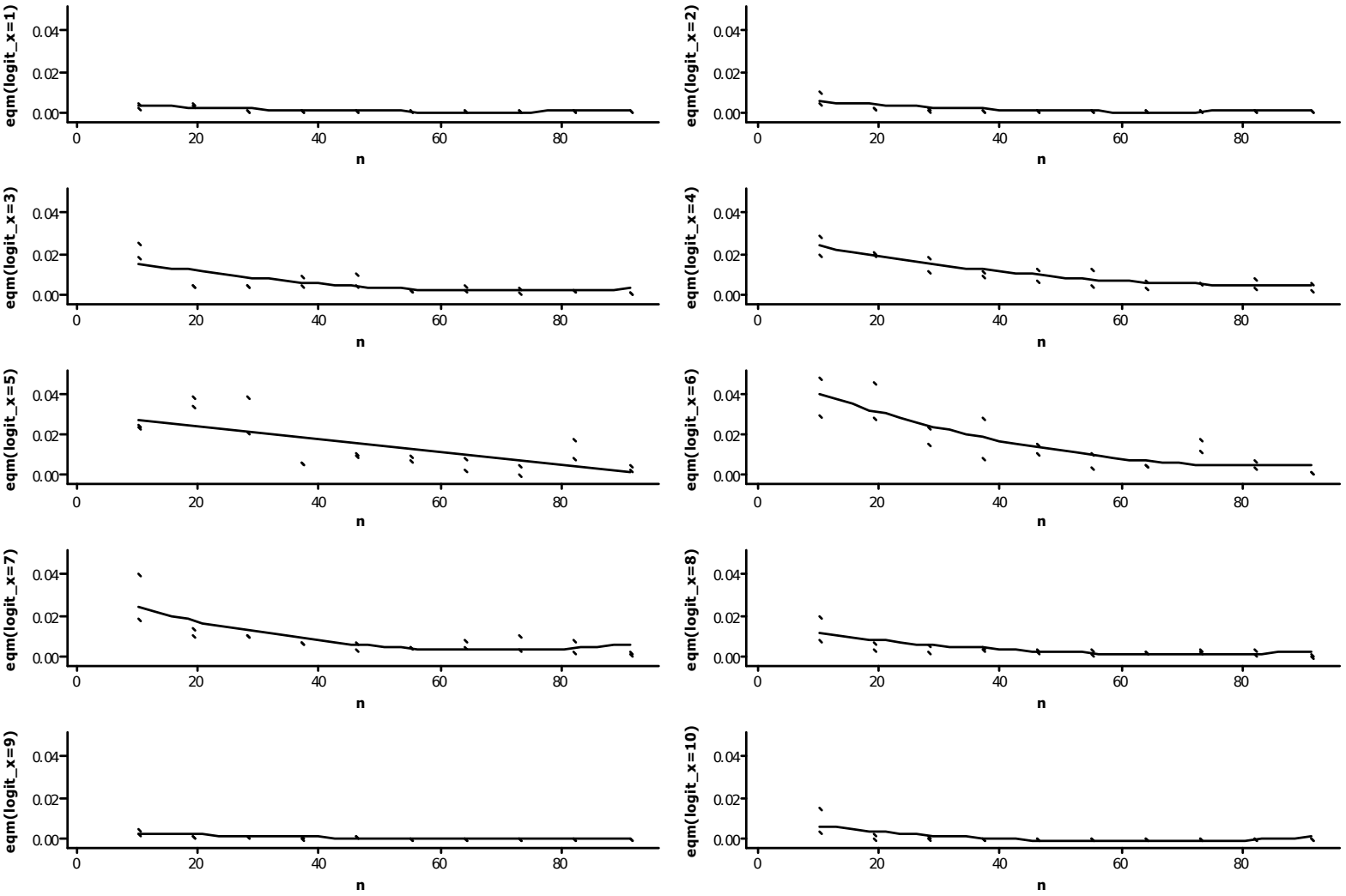

Figura 5 – Erro quadrático médio da probabilidade dos dados oriundos do logit fixados $x = 1,...,10$

Para os dados simulados a partir da função de ligação probit fixados diferentes níveis de *x*, as equações de regressão ajustadas estão apresentadas na Tabela 4 e as curvas na Figura 6.

Tabela 4 – Equação de regressão e grau de ajustamento para o erro quadrático médio 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	$E\hat{Q}M$ = 0,000134 - 0,000004 *n	0,29
x=2	$E\hat{Q}M$ = 0,000835 - 0,000030 *n	0,39
x=3	$E\hat{Q}M$ = 0,00715 - 0,000244* n + 0,000002* n^2	0,42
x=4	$E\hat{Q}M$ = 0,0302 - 0,000961* n + 0,000008 *n^2	0,71
x=5	$E\hat{Q}M$ = 0,0491 - 0,000833* n + 0,000005* n^2	0,80
x=6	$E\hat{Q}M$ = 0,0451 - 0,000571* n + 0,000002* n^2	0,54
x=7	$E\hat{Q}M$ = 0,0337 - 0,00106* n + 0,000008* n^2	0,44
x=8	$E\hat{Q}M$ = 0,00564 - 0,000172* n + 0,000001* n^2	0,47
x=9	$E\hat{Q}M$ = 0,000564 - 0,000018 *n	0,75
x=10	$E\hat{Q}M$ = 0,000160 - 0,000006 *n	0,69

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

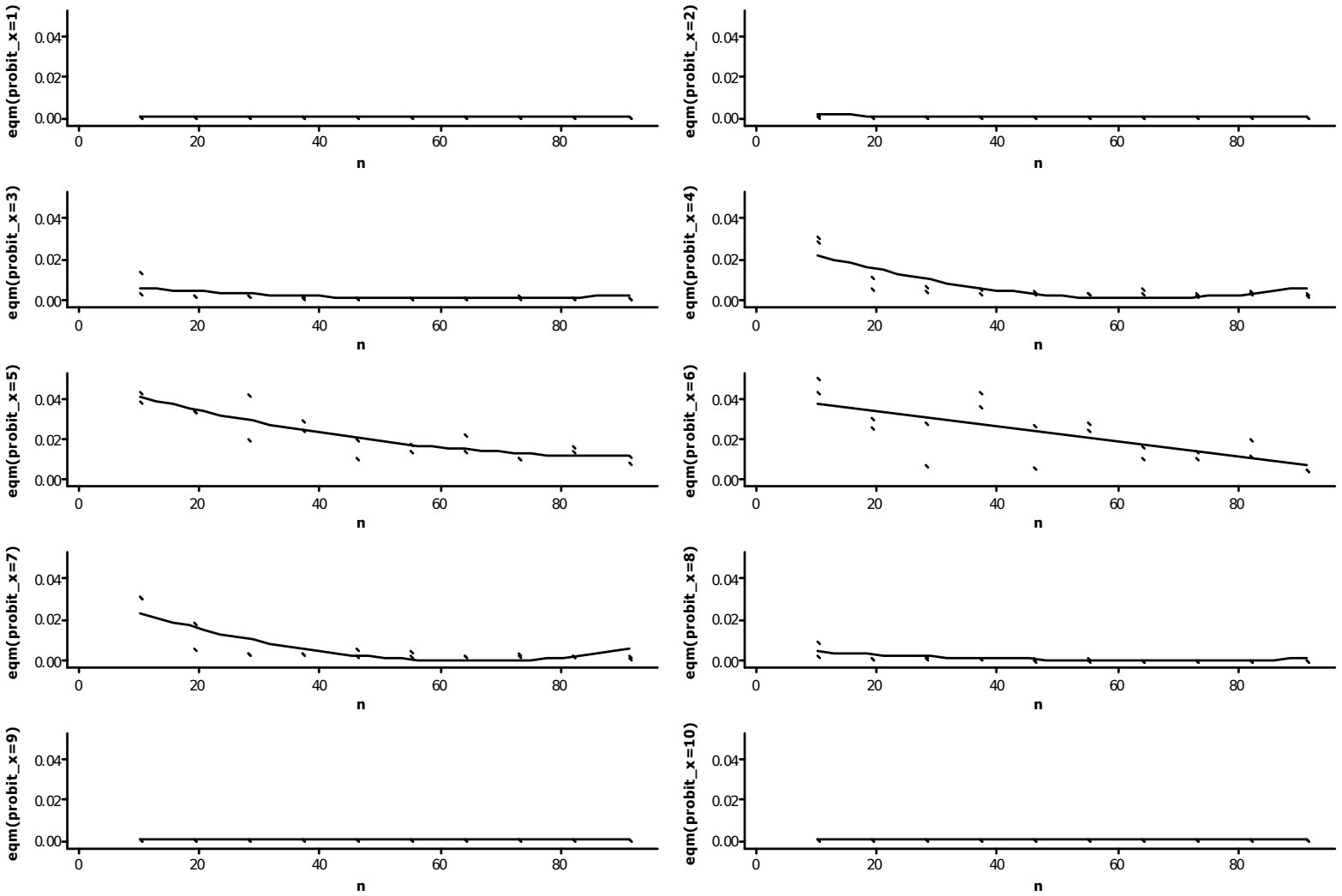

Figura 6 – Erro quadrático médio da probabilidade dos dados oriundos do probit fixados $x = 1,...,10$

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

3.4. Teste de Wald

Não houve diferença quanto à significância da estatística (*W*) do teste de Wald (P>0,05) dos dados que tiveram origem nas funções logit e probit como podem ser observados nas equações 13 a 18:

$$W\beta_0\_\text{logit} = -0{,}138 + 0{,}0189 * n - 0{,}000115 * n^2 \quad (R^2=0{,}982) \quad \textbf{[13]}$$

$$W\beta_1\_\text{logit} = -0{,}133 + 0{,}0212 * n - 0{,}000138 * n^2 \quad (R^2=0{,}965) \quad \textbf{[14]}$$

$$W\beta_0/\beta_1\_\text{logit} = -0{,}135 + 0{,}0183 * n - 0{,}000111 * n^2 \quad (R^2=0{,}983) \quad \textbf{[15]}$$

$$W\beta_0\_\text{probit} = -0{,}163 + 0{,}0233 * n - 0{,}000155 * n^2 \quad (R^2=0{,}978) \quad \textbf{[16]}$$

$$W\beta_1\_\text{probit} = -0{,}127 + 0{,}0244 * n - 0{,}000169 * n^2 \quad (R^2=0{,}938) \quad \textbf{[17]}$$

$$W\beta_0/\beta_1\_\text{probit} = -0{,}162 + 0{,}0229 * n - 0{,}000152 * n^2 \quad (R^2=0{,}979) \quad \textbf{[18]}$$

A significância do parâmetros e da constante aumenta em função do aumento de *n* (P<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 7).

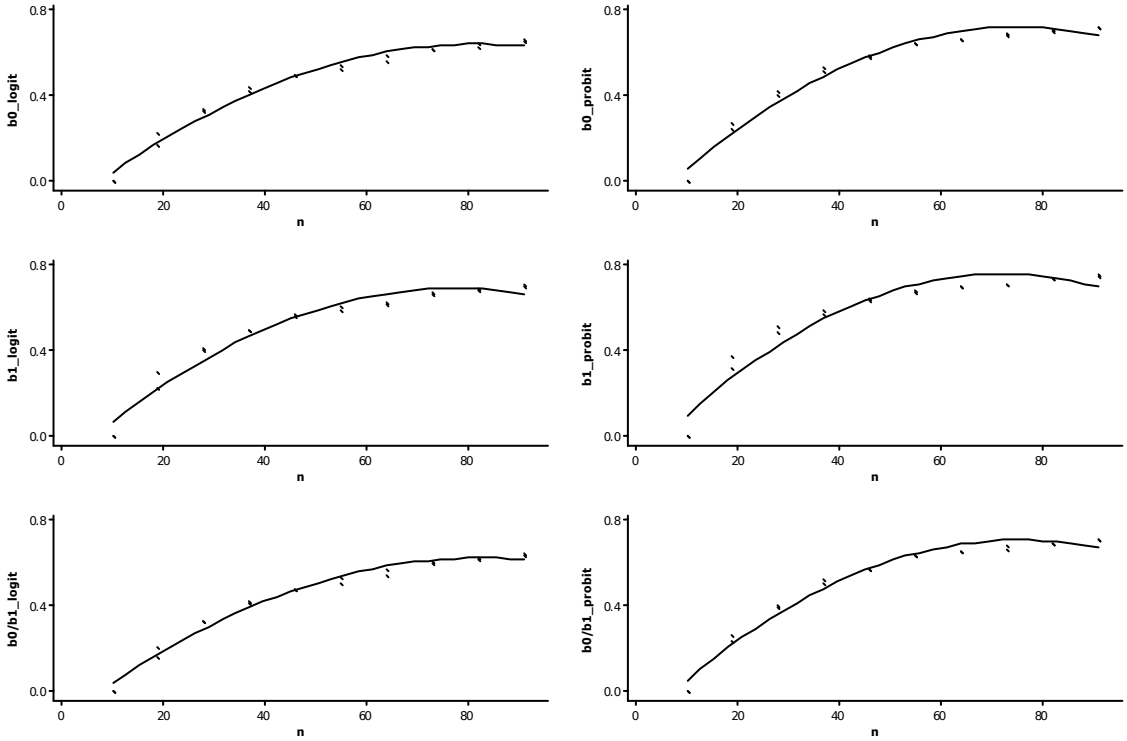

Figura 7 – Teste de Wald para a constante, coeficiente e constante com o coeficiente para os dados que tiveram origem nas funções de ligação logit e probit

De acordo com Queiroz (2011), o teste de Wald apresenta baixo desempenho em amostras pequenas; como pode ser observada na figura 7, a porcentagem de amostras em que os parâmetros foram significativos foi pequena para menores tamanhos de amostras. Ramalho e 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.

4. 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, isto é, o aumento do tamanho da amostra melhora a qualidade dos parâmetros de regressão binárias obtidas a partir das funções de ligação logit e probit.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. **Categorical data analysis**. New York: John Wiley & Sons, 1990.

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 jan. 2013.

BOLFARINE, H.; SANDOVAL, M. **Introdução à inferência estatística**. São Paulo: Coleção Matemática aplicada, Sociedade Brasileira de Matemática, 2000.

CASELLA, G.; BERGER, R. L. **Inferência estatística**. Tradução da 2ª edição Norte Americana: Solange Aparecida Visconde. São Paulo: Cengage Learning, 2010.

CORDEIRO, G.; DEMÉTRIO, C. **Modelos lineares generalizados**. In: Simpósio de estatística aplicada à experimentação agronômica – SEAGRO, 12.; reunião anual da região brasileira da sociedade internacional de biometria – RBRAS, 52., 2007, Santa Maria. **Minicurso.** Santa Maria: UFSM, 2007.

CORRAR, L. J.; PAULO, E.; FILHO, J. M. D. **Análise multivariada**: para os cursos de Administração, Ciências Contábeis e Economia. São Paulo: Atlas, 2009.

CRAMER J. S. **The origins and development of the logit model**. University of Amsterdam and Tinbergen Institute, Amsterdam, 2003. Disponível em: <http://www.cambridge.org/resources/0521815886/1208_default.pdf>. Acesso em: 22 jan. 2013.

DEMÉTRIO, C. G. P. **Modelos lineares generalizados em experimentação agronômica**. (Apostila) – Escola Superior de Agricultura Luiz de Queiroz Departamento de Ciências Exatas – LCE – USP, Piracicaba, SP, 2002. Disponível em: < http://ce.esalq.usp.br/clarice/Apostila.pdf>. Acesso em: 22 jan. 2013.

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

GUJARATI, D, N. **Econometria básica**. 3ª edição. São Paulo: Makron Books, 2005.

HAIR, J. F. J.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Análise multivariada de dados**. 6ª edição. Porto Alegre: Bookman, 2009.

LEMONTE, A. J. **Inferência sobre os parâmetros da distribuição Birnbauum-Saunders bi-paramétrica**. (Dissertação) – Universidade Federal de Pernambuco – UFPE, Recife, PE, 2006.

LIRA, S. A. **Efeitos do erro amostral nas estimativas dos parâmetros do modelo fatorial ortogonal.** (Tese) – Universidade Federal do Paraná – UFPR**,** Curitiba, PR, 2008.

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

MAGALHÃES, M. N.; LIMA, A. C. P. **Noções de probabilidade e estatística**. 6ª edição. São Paulo: Editora da Universidade de São Paulo, 2008.

MCCULLAGH, P.; NELDER, J.A. **Generalized linear models**. 2nd ed. Chapman & Hall/CRC, Boca Raton, Florida. 1989.

MIOT, H. A. Tamanho da amostra em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, v.10, p.275-278. 2011

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.

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, ISSN 0103-8478, 2011.

PENG, C. J.; SO, T. 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, n. 3, June 2002.

PEREIRA, J, M.;, DOMÍNGUEZ, M. A. C.; OCEJO, J. L. S. Modelos de previsão do fracasso empresarial: aspectos a considerar. **Revista de Estudos Politécnicos - Polytechnical Studies Review**. v. IV, n.7 111-148, ISSN: 1645-9911, 2007.

QUEIROZ, M. P. F. **Testes de hipóteses em regressão beta baseados em verossimilhança perfilada ajustada e em bootstrap**. (Dissertação) Universidade Federal de Pernambuco Centro de Ciências Exatas e da Natureza Departamento de Estatística. Pernambuco, 2011.

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, Version 2.13.0. Disponível em: http://www.R-project.org. Acesso em: 29 abr. 2012.

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.

RESENDE, M. D. V.; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Rev. Mat. Estat**., São Paulo, v. 20: p. 39-65, 2002.

STOCK, J. H.; WATSON, M. W. **Econometria**. São Paulo: Pearson Addison-Wesley, 2004.

APÊNDICE

Códigos de programação no software R

```
logit-logit
#############


Determinação da equaçao verdadeira
###################################

#set.seed=1234567
k=500  #número de amostras
n=60       #intervalo
probabilidade=NULL # vetor (o "NULL"cria um vetor de qualquer
tamanho)
############################################
#Gerando as amostra em forma de matrix(n,k)#
############################################
b0=-5.5
b1=1     #n=seq(10,100,10) #Tamanho da amostra n=10...100
x=seq(1,10,(10/n))
n=length(x)     #tamahno da amostra
y_logit=matrix(0,n,k) # matriz de zeros, com n linhas e k colunas
eta=b0 + b1*x   #Determinação da equaçao verdadeira#
logit=binomial(link="logit")$linkinv
for (j in 1:k)
{
y_logit[,j]=(rbinom(n,1, logit(eta)))
y_logit
}

####################################
#Colocando o x,y,amostra lado a lado#
####################################
y_aux=as.vector(y_logit)
amostra<-rep(1:k,each=n) #separaçao das amostras
x_aux=rep(x,n*k,n*k)
dados=cbind(y_aux,x_aux,amostra)
dados<-as.data.frame(dados)
############################################
#estimando os coeficientes b0,b1,convergencia#
############################################
coef_est=matrix(0,k,2)
converg=matrix(0,k,1)
erro_padrao_coef=matrix(0,k,2)
b0_aux=rep(b0,k*n)
dados_aux=as.data.frame(cbind(y_aux,x_aux,amostra,b0_aux))
for(i in 1:k)
{
coef_est[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "logit"))$coefficients
converg[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "logit"))$converged
erro_padrao_coef[i,]=summary(glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link =
"logit")))$coefficients[3:4]
}
```

```
Calculo da probabilidade geral
############################
fim=cbind(coef_est,converg)
nconv=length(fim[fim[,3]==1,3])
fim1=matrix(fim[fim[,3]==1],nconv,ncol(fim))  #se converge 1, nem
todas convergem
k1=nrow(fim1)
#probabilidade geral
prob_estimada=matrix(0,k1,n)
for(i in 1:k1)
{
for(j in 1:length(x))
{
prob_estimada[i,j]= logit(fim1[i,1]+fim[i,2]*matrix(x,1,n)[1,j])
}
}
prob_estimada1=cbind(matrix(t(prob_estimada),n*k1,1,byrow=T))
prob_obs=cbind(rep(logit(eta),k1))
EQM_pi=(sum((prob_obs-prob_estimada1)^2))/(k1*n)

#probabilidade especifica
x_esp=1
prob_fim=cbind(rep(x,k1),prob_obs, prob_estimada1)
prob_esp=cbind(rep(x,k1)==x_esp,prob_fim[,2]-prob_fim[,3])
prob_esp1=matrix(prob_esp[prob_esp[,1]==1],nconv)
EQM_esp=(sum((prob_esp1[,2])^2))/(k1*n)
EQM_esp


#Teste Wald#
############

#Erro padrão#
#############
erro_padrao_coef_1=cbind(erro_padrao_coef,converg)
erro_padrao_coef_CONVERG=erro_padrao_coef_1[which(erro_padrao_coef_1
[,3]==1),]

alpha=0.05
estatistica_wald=matrix(0,nrow(fim1),5)
colnames(estatistica_wald)=c("wald_b0","wald_b1","rej_b0","rej_b1","
Modelo_aceito")
#1=não rejeita
#0=rejeita
      for (j in 1:2){
            for (i in 1:nrow(fim1)) {
      estatistica_wald[i,j]=fim1[i,j]/erro_padrao_coef_CONVERG[i,j]
}}
            for (i in 1:nrow(fim1)) {
                        if(abs(estatistica_wald[i,1])>=qnorm(1-
alpha/2)) {estatistica_wald[i,3]=1}
                        if(abs(estatistica_wald[i,2])>=qnorm(1-
alpha/2)) {estatistica_wald[i,4]=1}
                        if (estatistica_wald[i,3]==1 &
estatistica_wald[i,4]==1){estatistica_wald[i,5]=1}
}
```

```
Porcentagem de modelos que ajustaram-se bem aos dados
#####################################################

percentual_b0=sum(estatistica_wald[,3])/nrow(estatistica_wald)
percentual_b1=sum(estatistica_wald[,4])/nrow(estatistica_wald)
percentual_b0_b1=sum(estatistica_wald[,5])/nrow(estatistica_wald)

logit-probit
#############

Determinação da equação verdadeira
##################################
set.seed=1234567
k=500  #número de amostras
n=10       #intervalo
probabilidade=NULL # vetor (o "NULL"cria um vetor de qualquer
tamanho)

Gerando as amostra em forma de matrix(n,k)
##########################################
b0=-5.5
b1=1     #n=seq(10,100,10) #Tamanho da amostra n=10...100
x=seq(1,10,(10/n))
n=length(x)     #tamahno da amostra
y_logit=matrix(0,n,k) # matriz de zeros, com n linhas e k colunas
eta=b0 + b1*x   #Determinação da equaçao verdadeira#
logit=binomial(link="logit")$linkinv
probit=binomial(link="probit")$linkinv
for (j in 1:k)
{
y_logit[,j]=(rbinom(n,1, logit(eta)))
y_logit
}

Colocando o x,y,amostra lado a lado
###################################

y_aux=as.vector(y_logit)
amostra<-rep(1:k,each=n) #separaçao das amostras
x_aux=rep(x,n*k,n*k)
dados=cbind(y_aux,x_aux,amostra)
dados<-as.data.frame(dados)

estimando os coeficientes b0,b1,convergencia
############################################
coef_est=matrix(0,k,2)
converg=matrix(0,k,1)
erro_padrao_coef=matrix(0,k,2)
b0_aux=rep(b0,k*n)
dados_aux=as.data.frame(cbind(y_aux,x_aux,amostra,b0_aux))
for(i in 1:k)
{
coef_est[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "probit"))$coefficients
converg[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "probit"))$converged
erro_padrao_coef[i,]=summary(glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link =
"probit")))$coefficients[3:4]
}
```

```
Calculo da probabilidade geral
################################

fim=cbind(coef_est,converg)
nconv=length(fim[fim[,3]==1,3])
fim1=matrix(fim[fim[,3]==1],nconv,ncol(fim))  #se converge 1, nem
todas convergem
k1=nrow(fim1)
#probabilidade geral
prob_estimada=matrix(0,k1,n)
for(i in 1:k1)
{
for(j in 1:length(x))
{
prob_estimada[i,j]= probit(fim1[i,1]+fim[i,2]*matrix(x,1,n)[1,j])
}
}
prob_estimada1=cbind(matrix(t(prob_estimada),n*k1,1,byrow=T))
prob_obs=cbind(rep(logit(eta),k1))
EQM_pi=(sum((prob_obs-prob_estimada1)^2))/(k1*n)

#probabilidade especifica
x_esp=1
prob_fim=cbind(rep(x,k1),prob_obs, prob_estimada1)
prob_esp=cbind(rep(x,k1)==x_esp,prob_fim[,2]-prob_fim[,3])
prob_esp1=matrix(prob_esp[prob_esp[,1]==1],nconv)
EQM_esp=(sum((prob_esp1[,2])^2))/(k1*n)


Teste Wald
############

Erro padrão
############
erro_padrao_coef_1=cbind(erro_padrao_coef,converg)
erro_padrao_coef_CONVERG=erro_padrao_coef_1[which(erro_padrao_coef_1
[,3]==1),]
alpha=0.05
estatistica_wald=matrix(0,nrow(fim1),5)
colnames(estatistica_wald)=c("wald_b0","wald_b1","rej_b0","rej_b1","
Modelo_aceito")
#1=não rejeita
#0=rejeita
      for (j in 1:2){
            for (i in 1:nrow(fim1)) {
      estatistica_wald[i,j]=fim1[i,j]/erro_padrao_coef_CONVERG[i,j]
}}
            for (i in 1:nrow(fim1)) {
                        if(abs(estatistica_wald[i,1])>=qnorm(1-
alpha/2)) {estatistica_wald[i,3]=1}
                        if(abs(estatistica_wald[i,2])>=qnorm(1-
alpha/2)) {estatistica_wald[i,4]=1}
                        if (estatistica_wald[i,3]==1 &
estatistica_wald[i,4]==1){estatistica_wald[i,5]=1}
}
```

```
Porcentagem de modelos que ajustaram-se bem aos dados
#####################################################
percentual_b0=sum(estatistica_wald[,3])/nrow(estatistica_wald)
percentual_b1=sum(estatistica_wald[,4])/nrow(estatistica_wald)
percentual_b0_b1=sum(estatistica_wald[,5])/nrow(estatistica_wald)

probit-logit
#############

Determinação da equação verdadeira
##################################
#set.seed=1234567
k=500  #número de amostras
n=10       #intervalo
probabilidade=NULL # vetor (o "NULL"cria um vetor de qualquer
tamanho)

Gerando as amostra em forma de matrix(n,k)
##########################################
b0=-5.5
b1=1     #n=seq(10,100,10) #Tamanho da amostra n=10...100
x=seq(1,10,(10/n))
n=length(x)     #tamahno da amostra
y_probit=matrix(0,n,k) # matriz de zeros, com n linhas e k colunas
eta=b0 + b1*x   #Determinação da equaçao verdadeira#
probit=binomial(link="probit")$linkinv
logit=binomial(link="logit")$linkinv
for (j in 1:k)
{
y_probit[,j]=(rbinom(n,1, probit(eta)))
y_probit
}


Colocando o x,y,amostra lado a lado
###################################
y_aux=as.vector(y_probit)
amostra<-rep(1:k,each=n) #separaçao das amostras
x_aux=rep(x,n*k,n*k)
dados=cbind(y_aux,x_aux,amostra)
dados<-as.data.frame(dados)

Estimação dos coeficientes b0,b1,convergência
#############################################
coef_est=matrix(0,k,2)
converg=matrix(0,k,1)
erro_padrao_coef=matrix(0,k,2)
b0_aux=rep(b0,k*n)
dados_aux=as.data.frame(cbind(y_aux,x_aux,amostra,b0_aux))
for(i in 1:k)
{
coef_est[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "logit"))$coefficients
converg[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "logit"))$converged
erro_padrao_coef[i,]=summary(glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link =
"logit")))$coefficients[3:4]
}
```

```
Calculo da probabilidade geral
##############################
fim=cbind(coef_est,converg)
nconv=length(fim[fim[,3]==1,3])
fim1=matrix(fim[fim[,3]==1],nconv,ncol(fim))  #se converge 1, nem
todas convergem
k1=nrow(fim1)
#probabilidade geral
prob_estimada=matrix(0,k1,n)
for(i in 1:k1)
{
for(j in 1:length(x))
{
prob_estimada[i,j]= logit(fim1[i,1]+fim[i,2]*matrix(x,1,n)[1,j])
}
}
prob_estimada1=cbind(matrix(t(prob_estimada),n*k1,1,byrow=T))
prob_obs=cbind(rep(logit(eta),k1))
EQM_pi=(sum((prob_obs-prob_estimada1)^2))/(k1*n)

#probabilidade especifica
x_esp=1
prob_fim=cbind(rep(x,k1),prob_obs, prob_estimada1)
prob_esp=cbind(rep(x,k1)==x_esp,prob_fim[,2]-prob_fim[,3])
prob_esp1=matrix(prob_esp[prob_esp[,1]==1],nconv)
EQM_esp=(sum((prob_esp1[,2])^2))/(k1*n)


Teste Wald
###########

Erro padrão
###########
erro_padrao_coef_1=cbind(erro_padrao_coef,converg)
erro_padrao_coef_CONVERG=erro_padrao_coef_1[which(erro_padrao_coef_1
[,3]==1),]

alpha=0.05
estatistica_wald=matrix(0,nrow(fim1),5)
colnames(estatistica_wald)=c("wald_b0","wald_b1","rej_b0","rej_b1","
Modelo_aceito")
#1=não rejeita
#0=rejeita
      for (j in 1:2){
            for (i in 1:nrow(fim1)) {
      estatistica_wald[i,j]=fim1[i,j]/erro_padrao_coef_CONVERG[i,j]
}}
            for (i in 1:nrow(fim1)) {
                        if(abs(estatistica_wald[i,1])>=qnorm(1-
alpha/2)) {estatistica_wald[i,3]=1}
                        if(abs(estatistica_wald[i,2])>=qnorm(1-
alpha/2)) {estatistica_wald[i,4]=1}
                        if (estatistica_wald[i,3]==1 &
estatistica_wald[i,4]==1){estatistica_wald[i,5]=1}
}
```

```
Porcentagem de modelos que ajustaram-se bem aos dados
####################################################
percentual_b0=sum(estatistica_wald[,3])/nrow(estatistica_wald)
percentual_b1=sum(estatistica_wald[,4])/nrow(estatistica_wald)
percentual_b0_b1=sum(estatistica_wald[,5])/nrow(estatistica_wald)


#probit-probit#
#############


#Determinação da equaçao verdadeira#
####################################
#set.seed=1234567
k=500  #número de amostras
n=100       #intervalo
probabilidade=NULL # vetor (o "NULL"cria um vetor de qualquer
tamanho)

#Gerando as amostra em forma de matrix(n,k)#
############################################
b0=-5.5
b1=1     #n=seq(10,100,10) #Tamanho da amostra n=10...100
x=seq(1,10,(10/n))
n=length(x)     #tamahno da amostra
y_probit=matrix(0,n,k) # matriz de zeros, com n linhas e k colunas
eta=b0 + b1*x   #Determinação da equaçao verdadeira#
probit=binomial(link="probit")$linkinv
for (j in 1:k)
{
y_probit[,j]=(rbinom(n,1, probit(eta)))
y_probit
}


#Colocando o x,y,amostra lado a lado#
####################################
y_aux=as.vector(y_probit)
amostra<-rep(1:k,each=n) #separaçao das amostras
x_aux=rep(x,n*k,n*k)
dados=cbind(y_aux,x_aux,amostra)
dados<-as.data.frame(dados)

Estimação dos coeficientes b0,b1,convergência
#############################################

coef_est=matrix(0,k,2)
converg=matrix(0,k,1)
erro_padrao_coef=matrix(0,k,2)
b0_aux=rep(b0,k*n)
dados_aux=as.data.frame(cbind(y_aux,x_aux,amostra,b0_aux))
for(i in 1:k)
{
coef_est[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "probit"))$coefficients
converg[i,]=glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link = "probit"))$converged
erro_padrao_coef[i,]=summary(glm(y_aux ~ x_aux, data =
dados[dados$amostra==i,],binomial(link =
"probit")))$coefficients[3:4]
}
```

```
Calculo da probabilidade geral
##############################
fim=cbind(coef_est,converg)
nconv=length(fim[fim[,3]==1,3])
fim1=matrix(fim[fim[,3]==1],nconv,ncol(fim))  #se converge 1, nem
todas convergem
k1=nrow(fim1)
#probabilidade geral
prob_estimada=matrix(0,k1,n)
for(i in 1:k1)
{
for(j in 1:length(x))
{
prob_estimada[i,j]= probit(fim1[i,1]+fim[i,2]*matrix(x,1,n)[1,j])
}
}
prob_estimada1=cbind(matrix(t(prob_estimada),n*k1,1,byrow=T))
prob_obs=cbind(rep(probit(eta),k1))
EQM_pi=(sum((prob_obs-prob_estimada1)^2))/(k1*n)

#probabilidade especifica
x_esp=1
prob_fim=cbind(rep(x,k1),prob_obs, prob_estimada1)
prob_esp=cbind(rep(x,k1)==x_esp,prob_fim[,2]-prob_fim[,3])
prob_esp1=matrix(prob_esp[prob_esp[,1]==1],nconv)
EQM_esp=(sum((prob_esp1[,2])^2))/(k1*n)
EQM_esp


Teste Wald
##########

Erro padrão
############
erro_padrao_coef_1=cbind(erro_padrao_coef,converg)
erro_padrao_coef_CONVERG=erro_padrao_coef_1[which(erro_padrao_coef_1
[,3]==1),]
alpha=0.05
estatistica_wald=matrix(0,nrow(fim1),5)
colnames(estatistica_wald)=c("wald_b0","wald_b1","rej_b0","rej_b1","
Modelo_aceito")
#1=não rejeita
#0=rejeita
      for (j in 1:2){
            for (i in 1:nrow(fim1)) {
      estatistica_wald[i,j]=fim1[i,j]/erro_padrao_coef_CONVERG[i,j]
}}
            for (i in 1:nrow(fim1)) {
                        if(abs(estatistica_wald[i,1])>=qnorm(1-
alpha/2)) {estatistica_wald[i,3]=1}
                        if(abs(estatistica_wald[i,2])>=qnorm(1-
alpha/2)) {estatistica_wald[i,4]=1}
                        if (estatistica_wald[i,3]==1 &
estatistica_wald[i,4]==1){estatistica_wald[i,5]=1}
}


Porcentagem de modelos que se ajustaram bem aos dados
#####################################################
percentual_b0=sum(estatistica_wald[,3])/nrow(estatistica_wald)
percentual_b1=sum(estatistica_wald[,4])/nrow(estatistica_wald)
percentual_b0_b1=sum(estatistica_wald[,5])/nrow(estatistica_wald)
```