MODELOS DE REGRESSÃO BINÁRIO COM ERRO DE CLASSIFICAÇÃO E ERRO DE MEDIDA

Patrícia de Souza Medeiros[1]; Betsabé Grimalda Blas Achic [2]

[1]Estudante do Curso de Estatística- CCEN – UFPE; E-mail: pdsm1@de.ufpe.br,
[2]Docente/pesquisador do Depto de Estatística – CCEN – UFPE. E-mail: betsabe@de.ufpe.br.

Sumário: Os modelos de regressão binário são aplicados em várias áreas de conhecimento, em especial na área da saúde. Neste trabalho apresentamos um estudo sobre modelos de regressão binário com erro de classificação e erro de medida. Estudos de simulação dos modelos de regressão binário com erros de classificação e de medida são desenvolvidos, e também uma aplicação é apresentada.

Palavras–chave: erro de classificação; erro de medida; regressão binária.

INTRODUÇÃO

Atualmente os modelos de regressão binário são um dos principais assuntos abordados entre as pesquisas realizadas no mundo acadêmico. Os modelos de regressão são utilizados em diversas áreas sendo elas, médica, epidemiológica, finanças, economia, etc. Dentre os livros e artigos que foram utilizados neste trabalho, nos baseamos principalmente em um estudo sobre modelos de regressão binário que incorporavam erro de classificação e erro de medida dado em Roy S, Banerjee T, Maiti T(2005), neste paper os autores consideraram as ligações *logito* e *probito*, e apenas apresentaram estudos de simulação para a ligação *probito*. Portanto nos sentimos motivados a estudar as propriedades utilizando a função de ligação *logito*. Os programas computacionais foram desenvolvidos no R-project e uma aplicação sobre os modelos de regressão binário foi apresentada.

MÉTODO

Este estudo foi dividido em algumas etapas, sendo elas,

1. Uma abordagem baseada em modelos de regressão para ajustar dados na presença de erros de medida em covariáveis e erro de classificação em respostas binárias.
2. Um modelo de regressão binário foi desenvolvido usando uma função de ligação que incorpora duas importantes fontes de erros, o erro de medida em covariáveis e o erro de classificação em respostas binárias.
3. O método de máxima verossimilhança foi usado para ajustar o modelo.
4. Uma aplicação foi realizada.

O MODELO DE REGRESSÃO BINÁRIO COM ERRO DE CLASSIFICAÇÃO E ERRO DE MEDIDA

Sejam *Y* a resposta binária não observada(verdadeira), e $\tilde{y}$ a resposta binária observada, $\mathbf{x}=(\mathbf{x}_1^{\mathrm{T}},\mathbf{x}_2^{\mathrm{T}})^{\mathrm{T}}$ em que $\mathbf{x}_1(p_1\times 1)$ e $\mathbf{x}_2(p_2\times 1);\ (p_1+p_2=p)$ são os preditores de interesse.

Assumimos que **X1** é observada sem erro de medida, e que **X2** não é observada, no entanto observa-se **Z**. Sejam ε_0 e ε_1 as probabilidades de cometer erro de classificação, ou seja,

$$\begin{aligned} P(\tilde{y}=1/y=0)&=\varepsilon_0\\ P(\tilde{y}=0/y=1)&=\varepsilon_1 \end{aligned} \qquad (1)$$

Para um valor fixo X assumimos que

$$P(y=1\,|\,\mathbf{x}) = g(\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{x}_2^{\mathrm{T}}\boldsymbol{\beta}_2) = g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \qquad (2)$$

Em que é uma $g^{-1}(\cdot)$ função de ligação apropriada, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$ é o parâmetro de regressão, β_0 é o primeiro componente de $\boldsymbol{\beta}_1$, e **1** é o primeiro componente de **X1**.
Também assumimos que os erros nas variáveis do modelo não são diferenciais, ou seja

$$f(y\,|\,\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) = f(y\,|\,\mathbf{x}_1, \mathbf{x}_2) \qquad (3)$$

Considerando os erros de medida assumimos que,

$$\mathbf{x}_2\,|\,\mathbf{z} \sim N_{p2}(h(\mathbf{z}), \Sigma_{22}(\mathbf{z})) \qquad (4)$$

em que $\Sigma_{22}(\mathbf{z})$ e $h(\mathbf{z}) = (h_1(\mathbf{z}), h_2(\mathbf{z}), \ldots, h_{p2}(\mathbf{z}))^{\mathrm{T}}$ são conhecidos. Assim, a partir de (1) e (2)

$$P(\tilde{y}=1\,|\,\mathbf{x}) = \varepsilon_0 + (1-\varepsilon_0-\varepsilon_1)g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) \qquad (5)$$

Considerando a equação **(3),**

$$P(\tilde{y}=1\,|\,\mathbf{x}_1, \mathbf{z}) = \varepsilon_0 + (1-\varepsilon_0-\varepsilon_1)\int g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) f(\mathbf{x}_2\,|\,\mathbf{z})\,\mathrm{d}\mathbf{x}_2$$

$$= \varepsilon_0 + (1-\varepsilon_0-\varepsilon_1)E_{\mathbf{x}_2|\mathbf{z}}\{g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\} \qquad (6)$$

em que $f(\mathbf{x}_2\,|\,\mathbf{z})$ é a função de densidade de probabilidade da distribuição normal dado em (4).
Note-se que (6) não é um modelo linear generalizado. Além disso, se $\varepsilon_0 + \varepsilon_1 = 1$, (6) não depende de $\boldsymbol{\beta}$, respostas manifestas não contêm qualquer informação sobre os parâmetros de regressão.

ESTIMAÇÃO DOS PARÂMETROS

Consideramos a estimativa dos parâmetros em dois casos:
(i) e são desconhecidos ;
(ii) As estimativas de e são e, e estão disponíveis a partir de um estudo de validação independente.

APLICAÇÃO, RESULTADOS E DISCUSSÃO

Durante o período de estudo desenvolvemos um programa computacional no pacote R-project que respeitasse todas as suposições apresentadas no tópico anterior. Este programa computacional realiza o mesmo procedimento utilizado por Roy S, Banerjee T, Maiti T(2005). Geramos amostras Monte Carlo e ajustamos os modelos descritos na seção anterior, seguindo o procedimento apresentado abaixo.

1. Geramos zi observado, com i = 1, ..., n, de uma uniforme (-4,4) e mantemos fixos.

2. Geramos xi de uma N (zi,σ^2), com i = 1, ..., n, para um valor σ^2=0.01 .
3. Geramos yi da distribuição de Bernoulli com probabilidade de sucesso g (β_o + $\beta_1$xi), com i = 1, ..., n, onde g é a função *logito*. Aqui utilizamos β_o= 0 e β_1= 1.
4. Geramos $\tilde{y}$ de yi usando P ($\tilde{y}$ = 1 | yi = 0) = e_0 e P ($\tilde{y}$ = 0 | yi = 1) = e_1 onde (e_0, e_1) são valores prefixados (0.05,0.05).
5. Tendo em conta os dados ($\tilde{y}$; zi), i = 1,..., n, utilizamos os modelos com erro de classificação e com erro de medida separadamente para estimar os parâmetros de interesse.

6. Repetimos os passos 2-5 para R=500.
7. Encontramos a média dos estimadores.

Nesta Tabela apresentamos as estimativas obtidas do estudo de simulação considerando a ligação *logito*. Também, podemos encontrar os resultados obtidos por Roy S, Banerjee T, Maiti T (2005), considerando ligação *probito*. Os resultados podem ser observados na Tabela 1.

Tabela 1. Resultado comparativo entre o uso das funções de ligação *Probito* e *Logito*, na obtenção das estimativas.

Estimador	**Modelo sem erro de classificação e sem erro de medida**	**Modelo com erro de classificação**	**Modelo com erro de medida**
Função de Ligação	***Logito*(desvio padrão) / *Probito*(desvio padrão)**	***Logito*(desvio padrão)/ *Probito*(desvio padrão)**	***Logito*(desvio padrão)/ *Probito*(desvio padrão)**
$\hat{\beta}_0$	0.003(0.027) / 0.004(0.015)	0.003(0.040) / 0.000(0.037)	-0.000(0.027) / 0.001(0.021)
$\hat{\beta}_1$	0.998(0.018) / 0.604(0.008)	1.003(0.039) / 0.997(0.046)	0.687(0.010) / 0.605(0.008)
$\hat{\varepsilon}_0$	-	0.065(0.007) / 0.050(0.006)	-
$\hat{\varepsilon}_1$	-	0.065(0.007) / 0.050(0.006)	-

Observamos na Tabela 1 que, em geral os desvios padrão de $\hat{\beta}_0$ e $\hat{\beta}_1$ correspondente a ligação *probito*, forneceram valores mais baixos comparados com os resultados de se considerar ligação *logito*. Nos dados dispostos na Tabela 1. é importante ressaltar que não foi incorporado simultaneamente erros de classificação e de medida ao modelo com erro de medida. Como aplicação, consideramos um exemplo apresentado em Hosmer D, Lemeshow S.(1989) para ajustar um modelo de regressão logística. Neste conjunto de dados existe uma relação entre a variável idade e a variável doença coronária, que é um tipo de doença cardíaca que causa um fornecimento inadequado de sangue ao músculo cardíaco – uma condição potencialmente danosa. A Tabela 2 lista a idade em anos (AGE) e presença ou ausência de evidência de doença coronária significativa (CHD) para as 100 pessoas selecionadas para participar deste estudo. A variável CHD é codificada com valor de zero para indicar CHD ausente, ou 1 para indicar CHD presente no indivíduo.

Tabela 2. Dados de idade(AGE) e evidências da doença cardíaca coronária(CHD) nos100 indivíduos entrevistados.

AGE	**CHD**
20 23 24 25 25 26 26 28 28 29 30 30 30 30 30 30 32 32 33 33 34 34 34 34 34 35 35 36 36 36 37 37 37 38 38 39 39 40 40 41 41 42 42 42 42 43 43 43 44 44 44 44 45 45 46 46 47 47 47 48 48 48 49 49 49 50 50 51 52 52 53 53 54 55 55 55 56 56 56 57 57 57 57 57 57 58 58 58 59 59 60 60 61 62 62 63 64 64 65 69	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 1 0 0 1 1 0 1 0 1 0 0 1 0 1 1 0 0 1 0 1 0 0 1 1 1 1 0 1 1 1 1 1 0 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 0 1 1 1

Para a aplicação dos modelos de regressão binário com erro de classificação e erro de medida, desenvolvemos um programa computacional no pacote R-project, para ajustar 3 tipos de modelos binários: o primeiro considera um modelo de regressão "sem erros" (Modelo sem erro de classificação e sem erro de medida); no segundo, os dados apresentam erro de classificação (Modelo com erro de classificação); já o terceiro apenas com a presença de erro de medida (Modelo com erro de medida). Ajustando estes três modelos nos dados da Tabela 2, obtemos os seguintes resultados.

Tabela 3. Resultados obtidos a partir do programa computacional desenvolvido.

Estimador	Modelo sem erro de classificação e sem erro de medida	Modelo com erro de classificação	Modelo com erro de medida
σ^2	-	-	0.01 / 0.04
$\hat{\beta}_0$	-5.31	-7.23	-5.31 / -5.31
$\hat{\beta}_1$	0.11	0.15	0.11 / 0.11
$\hat{\varepsilon}_0$	-	0.07	-
$\hat{\varepsilon}_1$	-	0.08	-

Sobre os resultados apresentados acima podemos destacar o fato do modelo com erro de medida apresentar estimativas iguais as estimativas do modelo sem erros, para σ^2=0.01, quanto para σ^2=0.04, neste conjunto de dados.

CONCLUSÕES

Neste trabalho, consideramos a modelagem da média de uma resposta binária assumindo uma função de ligação. No caso de ligação *probito* a expectativa sobre o lado direito de (6) pode ser avaliado de forma explícita, enquanto para ligação *logito* uma aproximação a este é obtido em (9). Para a análise nesse trabalho utilizamos a ligação logito.

Esperamos em trabalhos futuros aprofundar mais nossos estudos sobre modelos de regressão binário, desenvolvendo agora os modelos com os erros de classificação e de medida incorporados simultaneamente nos dados.

AGRADECIMENTOS

Agradeço primeiramente a Deus, a minha família, a professora Betsabé Achic que esteve me orientando durante toda a pesquisa e aos meus companheiros de curso, que me ajudaram bastante. Ao PIBIC/CNPq pelo apoio financeiro.

REFERÊNCIAS

Sposto R, Preston DL, Shimizu Y, Mabuchi K. The effect of diagnostic misclassification on non cancer andcancer mortality dose response in A-bomb survivors. Biometrics 1992; 48:605-617

Roy S, Banerjee T, Maiti T. Measurement error model for misclassified binary responses. Statistics in Medicine 2005; 24:269–283

Hosmer D, Lemeshow S. Appliied Logistic Regression. Wiley Interscience Publication. Massachusetts, 1989; 2-18.