Modelos de regressão com erro de medida aplicado a dados de laboratórios de ensaios.

Jaqueline Trentino SIlva [1], **Eveliny Barroso da Silva** [2]

1 Introdução

Sempre é interessante conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer, sobre outras. Mesmo que não exista relação causal entre as variáveis podemos relacioná-las por meio de uma expressão matemática, que pode ser útil para se estimar o valor de uma das variáveis quando conhecemos os valores das outras (estas de mais fácil obtenção ou antecessoras da primeira no tempo), sob determinadas condições. Genericamente, tais relações funcionais podem ser representadas por $Y = f(X_1, X_2, \ldots, X_n)$ onde Y representa a variável dependente e os $X_i, i = 1, \ldots, n$ representam as variáveis explanatórias (HOFFMANN *et al.*,2016).

Em situações práticas, pode acontecer que algumas covariáveis associadas á variável resposta sejam medidas com erro. Suponha que exista apenas uma covariável medida com erro, X_i, e que W_i seja a covariável realmente observada no modelo. Nesta situação, supomos que existe uma estrutura que relaciona a covariável observada W_i com a verdadeira covariável não observada X_i. Essa estrutura é, usualmente, aditiva, $W_i = X_i + \epsilon_i$, ou multiplicativa, $W_i = X_i\epsilon_i, i = 1, \ldots, n$ (SILVA, 2016). Para ilustrar a ideia, considere o relacionamento entre produtividade do milho (Y) e nitrogênio disponível no solo (X). Supondo que o modelo linear é uma aproximação adequada para esta situação, o coeficiente β_1 será o quanto a produtividade aumenta para o aumento de uma unidade de nitrogênio no solo. Para a estimativa de nitrogênio no solo será retirada uma amostra e uma análise feita no laboratório. No resultado da análise de laboratório não observa-se X_i e sim uma estimativa de X_i. Portanto, $W_i = X_i + \varepsilon_i$ será a variável observada nitrogênio, onde ε é o erro de medida resultante da amostragem e análise de laboratório.

Na literatura há várias propostas de trabalhos que abordam a teoria de modelos de regressão com erro de medida aditivo, dos quais podemos citar GUOLO and BRAZZALE 2008), PATRIOTA(2010), GUOLO(2011), SKRONDAL and KUHA(2012), TOMAYA *et al.*(2014), CASTRO and VIDAL(2017), RONDON *et al.*(2017) e FIGUEROA-´ZÚÑIGA *et al.*(2018).

Este trabalho foi motivado a partir de uma aplicação apresentada na dissertação de mestrado de MENDES(2017). O objetivo de MENDES(2017) foi avaliar os atributos microbiológicos do solo em áreas com diferentes níveis de produtividade de soja. Em seu trabalho, foram avaliadas 34 áreas de plantios comerciais cultivadas com soja no ano agrícola 2015/16, localizadas nos estados do Paraná, São Paulo, Minas Gerais, Goiás e Mato Grosso. As amostras foram coletadas na profundidade de 0 a 20 cm durante o período de florescimento pleno da cultura. Foram avaliados os teores de carbono orgânico total (COT), relação C:N, macro e micronutrientes, argila, carbono da biomassa microbiana, respiração basal do solo (RBS), quociente metabólico (qCO2) e microbiano (qMIC),

[1]Departamento de Estatística, Universidade Federal de Mato Grosso - UFMT. e-mail: *jaquetrentino@hotmail.com*

[2]Departamento de Estatística, Universidade Federal de Mato Grosso - UFMT. e-mail: *eveliny.stat@gmail.com*

hidrólise do diacetato de fluoresceína e a atividade das enzimas β-glucosidase, arilsulfatase e fosfatase ácida. Para cada variável em cada área foram feitas três medições. Essas medições caracterizam replicas de observações para cada variável e por esta razão pode envolver erro de medição além do erro de medida resultante da amostragem e análise de laboratório.

Portanto, neste trabalho é apresentado o que são os modelos de regressão com erro de medida nas covariáveis, as representações do erro de medida e alguns métodos de estimação que podem ser utilizados para corrigir o problema do erro de medida nas variáveis.

2 Metodologia

2.1 Modelos de Regressão com Erro de Medida nas covariáveis

Modelos de regressão são utilizados em diversas áreas com o intuito de relacionar uma variável resposta a variáveis explicativas (ou covariáveis) (RODRIGUES, 2013). Na prática, é comum que pelo menos uma variável explicativa não seja observada de forma exata, mas sim com algum tipo de erro de medida (CUNHA and COLOSIMO, 2003). Esses erros podem ocorrer devido a várias circunstâncias, das quais podemos citar:

- Erro de respostas em métodos de coleta de dados, como entrevistas ou questionários, causado por confusão, ignorância, por falta de cuidados, gerados por falta de treinamento adequado ou mesmo pelo método usado para obter a resposta.
- Erro de coleta dos dados por falha nos equipamentos devido ao desgastes dos componentes, falta de calibração ou a condições ambientais, que geram variabilidade em instrumento de leitura.
- Tempo e custo oriundos da observação da variável de interesse são inviáveis para o estudo.
- Processamento inadequado de dados ou perda de informações.
- Outros problemas que podem ocorrer ápos a coleta de dados.

Os modelos com erros nas variáveis são utilizados quando as covariáveis do modelo de regressão estão sujeitas a erros de medição. Na presença de uma covariável medida com erro em um modelo de regressão é extremamente importante determinar a relação existente entre a covariável observada e a covariável não observada. Os erros podem ser classificados de duas formas: erro aditivo ou erro multiplicativo.

2.2 Modelo

A construção desse modelo foi motivada pela aplicação retirada do trabalho de MENDES(2017). Considere um modelo de regressão no qual as variáveis respostas independentes $Y_1, \ldots, Y_n$, com suporte nos reais, estejam associadas a uma única covariável positiva medida sem erro Z_i, $i = 1, \ldots, n$ e a uma única covariável positiva medida com erro X_i, $i = 1, \ldots, n$. Com base no suporte da variável resposta, neste trabalho supomos que $(Y_i|X_i, Z_i) \sim N\left(\mu_i, \sigma_y^2\right)$. A função densidade para $Y_i|X_i, Z_i$ é dada por

$$f_{Y_i|X_i,Z_i}(y_i|x_i, z_i) = \frac{1}{\sigma_y\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma_y}\right)^2}, i = 1, \ldots, n. \quad (1)$$

Para relacionar os parâmetros da distribuição e os dados, usamos uma função de ligação linear,

$$\mu_i = \beta_0 + \beta_1 X_i + \gamma Z_i, \quad i = 1, \ldots, n,$$

em que $\beta_0 \in$ é o intercepto, $\beta_1 \in$ o coeficiente da covariável X_i medida com erro e γ representa o coeficiente da covariável Z_i medida sem erro, $i = 1, \ldots, n$.

Como há presença de replicas das variáveis, X_i será substituído por $\overline{X}_i$ e Y_i por $\overline{Y}_i$. Pela complexidade, optamos por trabalhar apenas com uma covariável (X_i) medida com erro (possui réplicas) e uma sem erro de medição Z_i (não possui réplicas). Apresentamos a seguir as principais representações dos erros de medida nas covariáveis e em seguida os métodos de estimação.

2.3 Representações do Erro de Medida

2.3.1 Erro Aditivo Clássico

O erro aditivo clássico considera que a covariável observada W_i é a soma da covariável não observada X_i e o erro de medida associado ϵ_i, ou seja,

$$W_i = X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

2.3.2 Erro Multiplicativo Clássico

O erro multiplicativo clássico considera que a covariável observada é o produto da covariável não observada pelo erro de medida associado, ou seja,

$$W_i = X_i \epsilon_i, \quad i = 1, \ldots, n.$$

2.4 Métodos de Estimação

"Em situações em que se deseja verificar a associação entre uma variável resposta e variáveis explicativas através de um modelo de regressão, os estimadores podem ser viciados se estas últimas estiverem sujeitas a erros de medição". Alguns estimadores foram propostos para corrigir ou reduzir o vício nessas situações. Entre os estimadores propostos estão o método de *Naive*, o método de calibração e o método de máxima verossimilhança (CUNHA and COLOSIMO, 2003).

2.4.1 Método *Naive*

O método *naive* ignora a presença do erro de medida no modelo e substitui a covariável não observada X_i pela covariável realmente observada $W_i, i = 1, \ldots, n$, ou seja,

$$\mu_i = \beta_0 + \beta_1 W_i + \gamma Z_i, \quad i = 1, \ldots, n. \tag{2}$$

Na presença de replicações da covariável W_i, substitui-se W_i por $\overline{W}_i, \quad i = 1, \ldots, n$. A principal conveniência em se utilizar esse método é que a regressão pode ser feita pelos métodos convencionais e a principal desvantagem é que é desconsiderada a presença do erro de medida na covariável o que pode gerar estimativas viciadas e inconsistentes (SILVA, 2018).

2.4.2 Método da Calibração

De acordo com SILVA(2018), o método calibração da regressão substitui a covariável não observável, X_i, pela estimativa de sua esperança condicionada a $W_i, E\left(X_i|W_i\right), i = 1, \ldots, n$, ou seja,

$$\mu_i = \beta_0 + \beta_1 E\left(X_i|W_i\right) + \gamma Z_i, \quad i = 1, \ldots, n. \tag{3}$$

em que, $E\left(X_i|W_i\right) = -w_i, \quad i = 1, \ldots, n$. Quando há replicas para $W_i, \quad i = 1, \ldots, n.$, substitui-se $w_i = \overline{w}_i, \quad i = 1, \ldots, n$.

2.4.3 Máxima Verossimilhança

Para a construção da função de verossimilhança, temos que determinar a distribuição conjunta de (Y_i, W_i, z_i). Como z_i é fixo e conhecido, não iremos considerá-lo na notação. A função densidade de (Y_i, W_i) pode ser obtida a partir da integração da função densidade conjunta dos dados completos (Y_i, W_i, X_i), denotada por $f_{Y_i,W_i,X_i}\left(y_i, w_i, x_i\right)$ com respeito a X_i, ou seja,

$$f_{Y_i,W_i}\left(y_i, w_i\right) = \int_{\Omega_{X_i}} f_{Y_i,W_i|X_i}\left(y_i, w_i|x_i\right) f_{x_i}\left(x_i\right) dx_i. \tag{4}$$

Para determinar a função densidade conjunta usamos a função acumulada para obter $f_{Y_i,W_i|X_i}\left(y_i, w_i|x_i\right)$.

$$F_{Y_i,W_i|X_i}\left(y_i, w_i|x_i\right) = P\left(Y_i \leq y_i, W_i \leq w_i|X_i = x_i\right), \quad i = 1, \ldots, n.$$

Considerando a suposição da presença de um erro de medida aditivo, substituímos W_i por $X_i + \epsilon_i$,como Y_i e ϵ_i são independentes, temos:

$$= F_{Y_i|X_i}\left(y_i|x_i\right) F_{\epsilon_i}\left(w_i - x_i\right) \quad i = 1, \ldots, n. \tag{5}$$

Derivando a equação (5) com respeito a $X_i, \quad i = 1, \ldots, n$, temos:

$$f_{Y_i,W_i|X_i}\left(y_i, w_i|x_i\right) = f_{Y_i|X_i,Z_i}\left(y_i|x_i, z_i\right) f_{\epsilon_i}\left(w_i - x_i\right). \tag{6}$$

Para a construção da função de verossimilhança, supomos que $\left(Y_i|X_i, Z_i\right) \sim N\left(\mu_i, \sigma_y^2\right), X_i \sim N\left(\mu_x, \sigma_x^2\right)$ e $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$. A escolha dessas distribuições se justifica primeiramente pelo suporte das variáveis envolvidas e segundo porque levando em consideração a presença de um erro de medida aditivo, temos

$$E\left(W_i\right) = E\left(X_i + \epsilon_i\right) = E\left(X_i\right) + E\left(\epsilon_i\right) = \mu_x + 0 = \mu_x = E\left(X_i\right),$$

$$\text{Var}\left(W_i\right) = \text{Var}\left(X_i + \epsilon_i\right) = \text{Var}\left(X_i\right) + \text{Var}\left(\epsilon_i\right) = \sigma_x^2 + \sigma_\epsilon^2.$$

Utilizando as equações apresentadas em (4) e (6), a função de verossimilhança é dada por:

$$L\left(\beta, \gamma\right) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} f_{Y_i,W_i|X_i}\left(y_i, w_i|x_i\right) \frac{1}{\sigma_x\sqrt{2\pi}} \mathrm{e}^{-\left(\frac{x_i - \mu_x}{\sqrt{2}\sigma_x}\right)^2} dx_i, \tag{7}$$

em que $\beta = (\beta_0, \beta_1)$ representam o intercepto e o coeficiente da covariável medida com erro e γ representa o coeficiente da covariável medida sem erro. Para solucionar a equação (7),

realiza-se a seguinte mudança de variável: $u_i = \frac{x_i - \mu_x}{\sqrt{2}\sigma_x}$ o que resulta em $du_i = dx_i/(\sqrt{2}\sigma_x)$ e $x_i = u_i\sigma_x\sqrt{2} + \mu_x$. Portanto,

$$L(\beta, \gamma) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} f_{Y_i, W_i|X_i}(y_i, w_i|x_i) \frac{1}{\sigma_x\sqrt{2\pi}} \mathrm{e}^{-u_i^2} \sigma_x\sqrt{2} du_i. \tag{8}$$

A integral exposta na equação (8) é bem complexa para se resolver analiticamente. Uma alternativa é resolvê-la computacionalmente utilizando o método da quadratura de Gauss-Hermite. Desta forma, a função log-verossimilhança pode ser escrita como,

$$l(\beta, \gamma) = \sum_{i=1}^{n} \log \left\{ \sum_{q=1}^{Q} \frac{p_q}{\sqrt{\pi}} f_{Y_i, W_i|X_{iq}}(y_i, w_i|x_{iq}) \right\}, \tag{9}$$

em que $x_{iq} = r_q\sigma_{x^*}\sqrt{2} + \mu_{x^*}$, Q é o número de pontos da quadratura, $r_1, \ldots, r_Q$ são as raízes do polinômio ortogonal de Hermite $H_q(x)$ (ou nós da quadratura) e $p_1, \ldots, p_Q$ são os pesos $p_q = \frac{2^{Q-1} n! \sqrt{\pi}}{n^2 [H_{Q-1}(uq)^2]}, q = 1, \ldots, Q$.

3 Aplicação a Dados Reais

Nesta seção apresentamos os resultados das estimativas obtidas utilizando o conjunto de dados reais obtidos através da dissertação de mestrado de MENDES(2017), da área de agricultura tropical. O interesse principal é verificar e efeito das variáveis β-glucosidase e a variável pH CaCl presente no solo com relação a produtividade de soja por hectare. A variável β-glucosidase foi medida três vezes, portanto consideramos que esta variável está sujeita a erro de medição além do erro de medida resultante da amostragem e análise de laboratório. No trabalho de MENDES(2017) outras variáveis foram coletadas mas devido a complexidade de um modelo com erro de medida consideramos para este estudo inicial apenas estas duas. Para cada variável em cada área foram feitas três medições. Essas medições caracterizam replicas de observações para cada variável e por esta razão pode envolver erro de medição além do erro de medida resultante da amostragem e análise de laboratório. Devido a complexidade, o modelo foi ajustado usando somente duas variáveis, a atividade da enzima β-glucosidase e com o pH CaCl presente no solo que é a variável sem erro de medida, pois nas três medições encontrou o mesmo resultado. O estudo consiste na avaliação de 34 áreas de plantios comerciais cultivadas com soja no ano agrícola 2015/16, localizadas nos estados do Paraná, São Paulo, Minas Gerais, Goiás e Mato Grosso. As amostras foram coletadas na profundidade de 0 a 20 cm durante o período de florescimento pleno da cultura. Se as variáveis preditoras fossem todas observadas de maneira exata, ou seja, se tivesse valores iguais em todas as três medições, poderíamos usar o modelo de regressão linear usual (RODRIGUES, 2013). O modelo foi ajustado via R pelos métodos *Naive* e Calibração e Máxima verossimilhança. Para estes métodos, X_i representa a variável medida com erro, β-glucosidase e a variável pH CaCl presente no solo representa a variável medida sem erro, Z_i.

Na Tabela (1) estão as estimativas dos parâmetros associados a cada covariável do modelo calculadas pelos métodos de estimação *naive*, calibração e máxima verossimilhança. Para os métodos *naive* e calibração, as estimativas dos parâmetros foram obtidas a partir da maximização do modelo (1). Em que para o método *naive*, usa-se a função de ligação (2) e para o método Calibração usa-se a função de ligação (3). Para o método de máxima

verossimilhança, as estimativas dos parâmetros foram obtidas a partir da maximização do modelo (9). As estimativas dos parâmetros associados às covariáveis sem erros de medida z_i não variam para nenhum método de estimação. Já as estimativas de β_1, parâmetro relativo a covariável com erro de medida sofrem uma pequena alteração.

Tabela 1: Estimativas dos parâmetros do modelo (1).

Método	Coeficientes		
	β_0	β_1	γ
Naive	1,9615	1,0318	1,0104
Calibração	1,9615	0,9691	1,0104
Máxima Verossimilhança	6,8911	0,0859	6,1833

4 Conclusões e Propostas Futuras

Neste trabalho temos por objetivo comparar os métodos de estimação que corrigem a presença do erro de medida nos dados e verificar qual o melhor.

Esta comparação será feita via estudo de simulação e análise de diagnóstico. Inicialmente testamos apenas uma covariavel para o banco de dados reais.

A próxima ação será realizar o estudo de simulação e análise de diagnóstico dos dados reais considerando apenas uma covariavel no modelo e posteriormente iremos acrescentar as demais que o pesquisador que cedeu os dados considere importante.

Agradecimentos

Agradeço à William Mendes por ter cedido os dados para este estudo.

Referencias Bibliográficas

CASTRO, M.; VIDAL, I. Bayesian inference in measurement error models from objective priors for the bivariate normal distribution. *Statistical Papers*, pages 120, 2017.

CUNHA, W. J.; COLOSIMO, E. A. Intervalos de confiança bootstrap para modelos de regressão com erros de medida. Rev. Mat. Estat, 2003. p. 2541.

FIGUEROA-ZÚÑIGA, J.; CARRASCO, J. M.; ARELLANO-VALLE, R.; FERRARI, S. L.; *et al.* A bayesian approach to errors-in-variables beta regression. *Brazilian Journal of Probability and Statistics*, 2018. p. 559582.

GUOLO, A. Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, 2011. p. 16391663.

GUOLO, A.; BRAZZALE, A. A simulation based comparison of techniques to corrent for measurement error in matched case-control studies. *Statistics in Medicine*, 2008. p. 37553775.

HOFFMANN, R; *et al.* Análise de regressão: uma introdução à econometria. O autor. 2016.

MENDES, W. M. Microbiologia dos solos cultivados em áreas com diferentes níveis de produtividade de soja. Dissertação (mestrado em agricultura tropical), 2017, Faculdade de Agronomia e Zootecnia da Universidade Federal de Mato Grosso. Acesso em: 2017-03-22.

PATRIOTA, A. G. Modelos heterocedásticos com erros nas variáveis. Ph.D. thesis, 2010, Universidade de São Paulo.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

RODRIGUES, A. S. Regressão logística com erro de medida: comparação de métodos de estimação. Dissertação (mestrado em estatística), 2013, IME-USP, São Paulo. Acesso em: 2014-01-15.

RONDON, L. M.; BOLFARINE, H.; *et al.* Bayesian analysis of flexible measurement error models. *Brazilian Journal of Probability and Statistics*, 2017. p. 618639.

SILVA, E. B. Contribuições em modelos de regressão com erro de medida multiplicativo. Tese (doutorado em estatística), 2016, IME-USP, São Paulo and Universidade Federal de São Carlos. Acesso em: 2018-02-22.

SILVA, E. B. D; DINIZ, C. A. R.; CARRASCO, J. M. F; CASTRO, M. D. A class of beta regression models with multiplicative log-normal measurement errors. *Communications in Statistics - Simulation and Computation*, 2018. p. 229248.

SKRONDAL, A.; KUHA, J. Improved regression calibration. *Psychometrika*, 2012. p. 649669.

TOMAYA, L. Y. C.; *et al.* Inferência em modelos de regressão com erros de medição sob enfoque estrutural para observações replicadas. 2014.