PEER REVIEW, Vol. 5, Nº 11, 2023
DOI: 10.53660/511.prw1508
ISSN: 1541-1389

Modelos com erros das variáveis: estudo de simulação e aplicação à dados de taxa de desemprego no Brasil

Measurement error models: simulation study and application to unemployment rate data in Brazil

Jaqueline Rodrigues de Souza Gentil[1], Tatiane Ferreira do Nascimento Melo[1]*, Tiago Moreira Vargas[1], Amanda Buosi Gazon Milani[1]

RESUMO

O modelo com erros nas variáveis é uma alternativa ao modelo clássico de regressão linear. Este modelo é utilizado quando as variáveis explicativas possuem erros inerentes à mensuração. Neste trabalho, realizamos estudos de simulação para analisar o desempenho do modelo com erro de medição, quando a suposição de normalidade é violada. Nesses casos, as estimativas dos parâmetros foram viesadas. Apresentamos ainda, uma aplicação a dados reais da taxa de desemprego, nos estados brasileiros, segundo o Índice de Gini, em 2019. Foram comparadas as regressões com e sem erros nas variáveis e selecionada aquela que melhor se ajustava aos dados. O modelo com erro nas variáveis explicou melhor o comportamento dos dados, sendo mais adequado para a análise de observações com erros de medida.

Palavras-chave: Erros de medida; Inferência; Modelo estrutural; Regressão simples;

ABSTRACT

The measurement error model is an alternative to the classic linear regression model. This model is used when the explanatory variables have errors inherent to the measurement. Here, we perform simulation studies to analyze the performance of the model with measurement error, when the assumption of normality is violated. In these cases, the parameter estimates were biased. We also present an application to real data of the unemployment rate in the Brazilian states according to the Gini Index in 2019. The regressions with and without errors in variables were compared and the one that best fit the data was selected. The errors-in-variables model better explained the behavior of the data, which is more appropriate for the analysis of observations with measurement errors.

Keywords: Measurement errors; Inference; Structural model; Simple regression.

1 Universidade Federal de Goiás
*E-mail: tmelo@ufg.br

Recebido: 18/04/2023 | Aceito: 20/05/2023 | Publicado: 12/06/2023

INTRODUÇÃO

Na análise de dados, muitas vezes estamos interessados em compreender a relação e comportamento entre variáveis. Para tanto, técnicas de regressão são amplamente utilizadas. Os procedimentos tradicionais de regressão assumem que os erros estão associados exclusivamente às variáveis dependentes e são decorrentes do ajuste do modelo (GILLARD, 2010). Contudo, na prática, há diversas situações em que as variáveis independentes podem apresentar erros inerentes ao processo de coleta das informações e às técnicas de medição (GILLARD, 2010). Estes casos são particularmente recorrentes nas ciências sociais e biológicas. É importante considerar que, quando há erros de medidas, substituir simplesmente as variáveis explicativas verdadeiras pelas observadas, em geral, leva a estimadores inconsistentes, denominados estimadores ingênuos (*naive*) (FULLER, 1987). Nesse contexto, um modelo que considera os desvios das variáveis independentes pode ser mais adequado e confiável. É o caso da regressão com erro nas variáveis.

Há décadas, a regressão com erro nas variáveis é aplicada na análise de dados. No decorrer do tempo, novas tecnologias têm sido incorporadas na compreensão desse modelo. Verificamos que estudos nacionais, como o de Carvalho Júnior et al. (2007), apropriam-se dessa técnica. Em sua pesquisa, os autores usaram o modelo aditivo estrutural e o combinaram com a construção de gráficos de controle de regressão. Desta forma, foi possível verificar que, independente do erro padrão utilizado (ajustado, condicional ou sobre valor predito), a apresentação gráfica é capaz de monitorar o modelo. A nível acadêmico, a pesquisa de Baba (2018) abordou a regressão com erro nas variáveis por meio da inferência bayesiana para estimação do coeficiente de inclinação do modelo. Por sua vez, De Oliveira et al. (2010) através da simulação de Monte Carlo verificou que os métodos *Plug-in* e Atenuador de Vício foram acurados somente para distribuições assimétricas apesar da elevada precisão dessas técnicas. Seguindo a linha temporal, Carrasco (2012) usou a regressão beta com erro de medidas. A estimação por máxima verossimilhança aproximada, máxima pseudo-verossimilhança aproximada e calibração da regressão também foram aplicadas a dados de simulação. Os mesmos métodos (verossimilhança e pseudoverossimilhança) foram comparados por Tomaya (2014) no modelo para observações replicadas. Em cada situação, analisou-se as propriedades dos estimadores. Finalmente, apresentamos o estudo de Soares (2020), com a abordagem via mistura finita para modelos de regressão linear com erro nas variáveis.

O trabalho propôs uma distribuição de mistura finita *skew*-Normal, com o objetivo de flexibilizar os erros independente da simetria dos dados.

De forma semelhante, podemos contextualizar as pesquisas internacionais sobre o tema. Al-Sharadqah et al. (2013) discutiram a presença de infinitos momentos na estimação por máxima verossimilhança da regressão circular e elíptica com erro nas variáveis. A característica da distribuição das variáveis também foi objeto de estudo de Nghiem et al. (2020). Estes autores propuseram a estimação para distribuições e variâncias desconhecidas. A regressão com erro nas variáveis foi ainda debatido por Reilly e Patino-Leal (1981), Thompson e Willis (1986), Bickel e Ritov (1987), Leamer (1987), Cheng e Van Ness (1991) e Gillard (2010).

Neste trabalho, propomos verificar o desempenho do modelo com erros nas variáveis quando o pressuposto de normalidade da técnica é violado. Levantamos a hipótese de que a regressão com erros nas variáveis pode ser utilizada para dados com distribuição *t*-Student quando o tamanho amostral é suficientemente grande. Além disso, aplicamos o modelo para a compreensão de dados reais de desemprego de acordo com valores de Índice de Gini. Este índice avalia a desigualdade econômica por meio da proporção de renda acumulada do indivíduo em relação à população. Ambas variáveis estão sujeitas a erros na coleta e no processamento inadequado dos dados.

O estudo foi planejado em três etapas. Primeiramente, discutimos o modelo e a estrutura da regressão com erros nas variáveis. Em seguida, comparamos o modelo utilizando diferentes distribuições das variáveis, assim como tamanhos amostrais. Finalmente, abordamos as características dos dados de desemprego e Índice de Gini e analisamos o desempenho da regressão em questão.

REGRESSÃO COM ERROS NAS VARIÁVEIS

No conceituado livro de Fuller (1987), ‘‘*Measurement error models’’*, são apresentados os aspectos fundamentais do modelo com erros nas variáveis. A abordagem dos erros de mensuração pode ser realizado sob três perspectivas. Na primeira, denominada modelo funcional, os valores conhecidos das variáveis independentes (x) são considerados parâmetros. Assim, cada observação (x_i) é uma constante desconhecida e o número de parâmetros cresce de acordo com o tamanho amostral. Por sua vez, no modelo estrutural assume-se uma distribuição de probabilidade para as variáveis explicativas não-observadas. No caso de uma variável $x_i$aleatória e com distribuição

Normal, temos média μ_x e variância σ_x^2. Finalmente, o modelo ultraestrutural é uma generalização dos modelos anteriores. Isto é, os $x_i's$ são variáveis aleatórias independentes, mas não identicamente distribuídas, podendo ter diferentes médias μ_{x_i} e variâncias $\sigma_{x_i}^2, \quad i$. Se $\mu_{x_1} = \cdots = \mu_{x_n} = \mu_x$, o modelo ultraestrutural se reduz ao modelo estrutural e, se $\sigma_x = 0$, o modelo ultraestrutural equivale ao modelo funcional (DOLBY, 1976). Em nosso estudo, focamos na abordagem estrutural com ênfase no pressuposto de distribuição Normal das variáveis resposta e explicativas. Independentemente da técnica adotada, novos métodos inferenciais, diferentes dos utilizados na análise de regressão simples, devem ser aplicados na regressão com erros nas variáveis.

Considere um modelo de regressão linear simples entre duas variáveis:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n. \tag{1}$$

Vamos considerar que a variável x_i é obtida com erro de mensuração devido às incertezas no momento da coleta dos dados, por exemplo, quando elas são observadas indiretamente. Denominaremos o erro da variável independente de u_i. Portanto, o seu valor observado é dado por:

$$X_i = x_i + u_i, i = 1, \dots, n. \tag{2}$$

As suposições do modelo são:

(i)$E(u_i) = E(\varepsilon_i) = 0, E(x_i) = \mu_i$;

(ii)$Var(u_i) = \sigma_{uu}, Var(x_i) = \sigma_{xx}, Var(\varepsilon_i) = \sigma_{\varepsilon\varepsilon}, \ i$;

(iii)$Cov(u_i, u_j) = Cov(x_i, x_j) = Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$;

(iv)$Cov(u_i, \varepsilon_j) = Cov(x_i, \varepsilon_j) = Cov(u_i, x_j) = 0, \quad i \neq j$.

Reescrevendo o modelo (1)-(2), matricialmente, temos que:

$$Z_i = \delta + \Delta b_i,$$

com $i = 1, \dots, n$, em que

$$Z_i = \begin{pmatrix} Y_i \\ X_i \end{pmatrix}, \quad \delta = \begin{pmatrix} \beta_0 \\ 0 \end{pmatrix}, \ \Delta = \begin{pmatrix} \beta_1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad b_i = \begin{pmatrix} x_i \\ \varepsilon_i \\ u_i \end{pmatrix}.$$

Assumimos que os erros $b_1, b_2, \dots, b_n$ são independentes e cada b_i tem distribuição Normal trivariada, ou seja:

$$b_i = \begin{pmatrix} x_i \\ \varepsilon_i \\ u_i \end{pmatrix} \sim N_3 \left(\begin{bmatrix} \mu_x \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{\varepsilon\varepsilon} & 0 \\ 0 & 0 & \sigma_{uu} \end{bmatrix} \right).$$

Logo, o vetor aleatório $Z_i = (Y_i, X_i)^T$, tem uma distribuição bivariada Normal, cujos momentos populacionais são expressos por:

$$Z_i = \begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N_2\left(\begin{bmatrix} \beta_0 + \beta_1\mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \beta_1^2\sigma_{xx} + \sigma_{\varepsilon\varepsilon} & \beta_1\sigma_{xx} \\ \beta_1\sigma_{xx} & \sigma_{xx} + \sigma_{uu} \end{bmatrix}\right).$$

Aqui, podemos compreender a diferença da regressão linear simples para a regressão com erro nas variáveis. Essa abordagem é usada para obter estimadores consistentes dos parâmetros de interesse. Para a regressão linear clássica, o método dos mínimos quadrados é normalmente utilizado. Contudo, medidas com erro nas variáveis independentes (x_i) são mais complexas. Na literatura, existem diversas técnicas para estimar os parâmetros no modelo com erros nas variáveis. Detalhes sobre a estimação por mínimos quadrados podem ser vistos em Fuller (1987), seção 1.3.3. Neste trabalho, usamos o método dos momentos para obter os estimadores dos parâmetros. Sob a suposição de normalidade os estimadores de método dos momentos são os estimadores de máxima verossimilhança ajustados para graus de liberdade. Ou seja, os estimadores de σ_{xx} e σ_{uu} diferem dos estimadores de máxima verossimilhança por um fator de $n(n-1)^{-1}$ (FULLER, 1987, pág. 31).

Para a aplicação do método dos momentos assumimos que os momentos de primeira e segunda ordem existem. Os momentos amostrais de primeira ordem são:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{e} \quad \overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i.$$

Os momentos amostrais de segunda ordem são denotados por:

$$S_{XX} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}, \quad S_{YY} = \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}, \quad \text{e} \quad S_{XY} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}.$$

Por sua vez, os momentos populacionais de primeira ordem são representados por:

$$\mu_X = \hat{\mu_x} \text{ e } \mu_Y = \hat{\beta_0} + \hat{\beta_1}\hat{\mu_x}.$$

Temos, ainda, os momentos populacionais de segunda ordem:

$$\sigma_{XX} = \hat{\sigma}_{xx} + \sigma_{uu}, \ \sigma_{XY} = \hat{\beta_1}\hat{\sigma}_{xx} \text{ e } \sigma_{YY} = \hat{\beta_1}^2\hat{\sigma}_{xx} + \sigma_{\varepsilon\varepsilon}.$$

Vale salientar que $E(X) = \mu_X, Var(X) = \sigma_{XX}$ e $Cov(X,Y) = \sigma_{XY}$. Estas informações nos permitirão determinar a média amostral de Z_i, que é dada por $\overline{Z} = \left(\overline{Y}, \overline{X}\right)^T$, e a sua matriz de covariância amostral

$$S_{ZZ} = (n-1)^{-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^T\left(Z_i - \overline{Z}\right),$$

especificada de acordo com os valores de S_{YY}, S_{XY} e S_{XX}.

Ao igualar os momentos amostrais e populacionais, isolamos os parâmetros desejados e encontramos os estimadores $\hat{\beta_0}, \hat{\beta_1}, \hat{\mu_x}, \hat{\sigma}_{xx}$ e $\hat{\sigma}_{\varepsilon\varepsilon}$. Isto é,

$$\hat{\beta_0} = \overline{Y} - \hat{\beta_1}\overline{X}, \ \hat{\beta_1} = \frac{S_{XY}}{S_{XX}-\sigma_{uu}}, \ \hat{\mu_x} = \overline{X}, \tag{3}$$

$$\hat{\sigma}_{xx} = S_{XX} - \sigma_{uu} \text{ e } \hat{\sigma}_{\varepsilon\varepsilon} = S_{YY} - \hat{\beta}_1 S_{XY}. \quad (4)$$

Note que, o erro de mensuração u_i tem variância desconhecida σ_{uu}, o que faz com que, das equações (3)-(4), tenhamos um problema de identificabilidade. Não é possível utilizar apenas os dados apresentados para determinar o modelo com erros nas variáveis porque temos somente cinco equações para estimar seis parâmetros. Essa dificuldade pode ser resolvida adicionando informações sobre o modelo. Desta forma, iremos considerar que a razão entre as variâncias dos erros u_i e ε_i é conhecida e determinada de acordo com $\lambda = \sigma_{\varepsilon\varepsilon}/\sigma_{uu}$. Como $\sigma_{\varepsilon\varepsilon} = \lambda\sigma_{uu}$, podemos redefinir a matriz de covariância amostral de Z_i como:

$$Z_i = \begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N_2\left(\begin{bmatrix} \beta_0 + \beta_1\mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \beta_1^2\sigma_{xx} + \lambda\sigma_{uu} & \beta_1\sigma_{xx} \\ \beta_1\sigma_{xx} & \sigma_{xx} + \sigma_{uu} \end{bmatrix}\right). \quad (5)$$

Neste caso, o vetor de parâmetros desconhecidos é dado por $\theta = (\beta_0, \beta_1, \mu_x, \sigma_{xx}, \sigma_{uu})^T$.

Sabemos que na regressão com erros nas variáveis $S_{XY} \neq 0$. Portanto, usando a matriz de covariâncias em (5), vamos resolver o sistema de equações:

$$\begin{aligned} S_{XY} &= \hat{\beta}_1\hat{\sigma}_{xx} \\ S_{YY} - \lambda S_{XX} &= \hat{\beta}_1^2\hat{\sigma}_{xx} - \lambda\hat{\sigma}_{xx} \\ \hat{\beta}_1^2 S_{XY} - \hat{\beta}_1(S_{YY} - \lambda S_{XX}) - \lambda S_{XY} &= 0 \end{aligned} \quad (6)$$

Portanto, do sistema (6) resultam os respectivos estimadores (FULLER, 1987):

$$\begin{aligned} \hat{\beta}_1 = \frac{S_{YY} - \lambda S_{XX} + \left[(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2\right]^{1/2}}{2S_{XY}}, \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}, \hat{\mu}_x = \overline{X}, \\ \hat{\sigma}_{xx} = \frac{\left[(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2\right]^{1/2} - (S_{YY} - \lambda S_{XX})}{2\lambda}, \\ \hat{\sigma}_{uu} = \frac{S_{YY} + \lambda S_{XX} - \left[(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2\right]^{1/2}}{2\lambda}. \end{aligned} \quad (7)$$

RESULTADOS NUMÉRICOS

Nesta seção, o objetivo é analisar o desempenho da regressão com erros nas variáveis quando o pressuposto de normalidade da distribuição das variáveis independentes e variável dependente é violado. Para tanto, levantamos a hipótese de que o modelo pode ser utilizado para dados com distribuição *t*-Student quando o tamanho amostral é suficientemente grande.

Diversos aspectos da inferência clássica estão fundamentados na normalidade dos dados. Contudo, para dados reais, muitas vezes nos deparamos com a presença de valores

extremos que podem afetar a estimação da média e variância. A análise de observações com erros de mensuração pode ser uma destas situações. Desta forma, estudar técnicas que permitam o pesquisador explorar dados com distribuição diferente da Normal pode representar uma vantagem no campo da aplicação da regressão linear.

Uma alternativa à distribuição Normal é a distribuição *t*-Student. Esta distribuição é simétrica e pertence à família das distribuições elípticas, assim como a Normal. Ela pode ser definida pelos parâmetros μ denominado vetor de locação, Σ, que corresponde à matriz de dispersão, e ν que representa os graus de liberdade da distribuição. Uma vantagem da distribuição *t*-Student é que possui caudas mais pesadas, acomodando melhor *outliers*. A função de distribuição da variável aleatória *t*-Student converge para a função de distribuição da variável aletaória Normal quando $\nu \to \infty$.

Retomando o contexto das simulações, as análises foram aplicadas ao modelo de regressão linear entre duas variáveis considerando que a variável x_i foi obtida com erro de mensuração, u_i, ou seja, o modelo (1)-(2). Assumimos que a razão entre as variâncias dos erros u_i e ε_i seja conhecida e determinada de acordo com $\lambda = \sigma_{\varepsilon\varepsilon}/\sigma_{uu}$. Desta forma, as simulações realizadas buscaram estimar o vetor de parâmetros $\theta = (\beta_0, \beta_1, \mu_x, \sigma_{xx}, \sigma_{uu})^T$ que define a distribuição conjunta Z_i dada em (5).

Neste estudo de simulações, os parâmetros foram fixados em $\theta = (0.5,1.0,5.0,1.5,0.5)^T$ e os estimadores dados em (7) foram utilizados. Utilizamos a técnica de Monte Carlo, com 10.000 réplicas. A variável Z_i, da equação (5), foi gerada usando a distribuição Normal ou *t*-Student com 5, 10, 20, 30 e 40 graus de liberdade. Para cada situação foram considerados tamanhos amostrais iguais a 20, 30, 40, 50, 100 e 200. Todas as análises foram realizadas com o auxílio do software R (R CORE TEAM, 2023).

Apresentamos, nas Tabelas 1 a 6, as estimativas dos parâmetros do modelo e seus respectivos vieses. Na Tabela 1, temos o caso em que os dados foram gerados a partir da distribuição Normal e para Tabelas 2 a 6 geramos os dados usando a distribuição *t*-Student com 5, 10, 20, 30 e 40 graus de liberdade, respectivamente.

Tabela 1 – Estimativas dos parâmetros e respectivos vieses quando usamos a distribuição Normal para gerar Z_i.

Parâmetros	n = 20		n = 30		n = 40	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,376	-0,124	0,425	-0,075	0,435	-0,065
β_1	1,025	0,025	1,015	0,015	1,013	0,013
μ_x	5,002	0,002	4,998	−0,002	4,999	−0,001
σ_{xx}	1,452	−0,048	1,470	−0,030	1,477	−0,023
σ_{uu}	0,445	−0,055	0,462	−0,038	0,473	−0,027
Parâmetros	n = 50		n = 100		n = 200	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,464	−0,036	0,483	−0,017	0,490	−0,010
β_1	1,008	0,008	1,004	0,004	1,002	0,002
μ_x	4,997	−0,003	4,999	−0,001	5,000	0,000
σ_{xx}	1,483	−0,017	1,491	−0,009	1,492	−0,008
σ_{uu}	0,478	−0,022	0,490	−0,010	0,495	−0,005

Fonte: Elaborada pelos autores

Observamos, através dos resultados numéricos, que para o modelo normal com erros nas variáveis, temos estimadores menos viesados quando os dados são gerados usando a distribuição Normal, como era esperado. Por exemplo, quando n = 20 e Z_i é gerada pela distribuição Normal, temos que $\hat{\beta}_0$ é 0,376 e seu viés -0,124, mas quando Z_i é gerada pela distribuição *t*-Student com 5 graus de liberdade temos que $\hat{\beta}_0$ é 0,312 e seu viés -0,188 (Tabelas 1 e 2).

Tabela 2 – Estimativas dos parâmetros e respectivos vieses quando usamos a distribuição *t*-Student para gerar Z_i, com $\nu = 5$ graus de liberdade.

Parâmetros	*n* = 20		*n* = 30		*n* = 40	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,312	−0,188	0,354	−0,146	0,385	−0,115
β_1	1,038	0,038	1,029	0,029	1,023	0,023
μ_x	5,004	0,004	4,996	−0,004	5,003	0,003
σ_{xx}	2,473	0,973	2,472	0,972	2,494	0,994
σ_{uu}	0,704	0,204	0,739	0,239	0,769	0,269
Parâmetros	*n* = 50		*n* = 100		*n* = 200	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,408	−0,092	0,455	−0,045	0,478	−0,022
β_1	1,019	0,019	1,009	0,009	1,004	0,004
μ_x	4,996	−0,004	4,999	−0,001	5,001	0,001
σ_{xx}	2,496	0,996	2,495	0,995	2,506	1,006
σ_{uu}	0,776	0,276	0,802	0,302	0,820	0,320

Fonte: Elaborada pelos autores

As estimativas de β_0 e β_1 convergem para os verdadeiros valores dos parâmetros conforme a amostra cresce, ou seja, quando aumentamos o tamanho amostral de $n = 20$ ($\hat{\beta}_1 = 1{,}038$ e viés = 0,038) para $n = 40$ ($\hat{\beta}_1 = 1{,}004$ e viés = 0,004) (Tabela 2). Esta situação se repete para o estimador de β_0. Isto reafirma a importância da utilização de um modelo normal quando os dados realmente provêm de uma população com distribuição Normal. Pelo Teorema Central do Limite, independentemente da distribuição da variável de interesse, a distribuição das médias amostrais tenderão a uma distribuição Normal à medida que o tamanho da amostra aumenta. Este resultado é evidenciado nestes estudos numéricos.

Tabela 3 – Estimativas dos parâmetros e respectivos vieses quando usamos a distribuição *t*-Student para gerar Z_i, com $\nu = 10$ graus de liberdade.

Parâmetros	*n* = 20		*n* = 30		*n* = 40	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,348	−0,152	0,395	−0,105	0,427	−0,073
β_1	1,029	0,029	1,022	0,022	1,015	0,015
μ_x	4,996	−0,004	5,001	0,001	4,999	−0,001
σ_{xx}	1,830	0,330	1,837	0,337	1,855	0,355
σ_{uu}	0,547	0,047	0,576	0,076	0,590	0,090
Parâmetros	*n* = 50		*n* = 100		*n* = 200	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,452	−0,048	0,474	−0,026	0,487	−0,013
β_1	1,010	0,010	1,005	0,005	1,002	0,002
μ_x	4,999	−0,001	5,000	0,000	5,002	0,002
σ_{xx}	1,863	0,363	1,869	0,369	1,872	0,372
σ_{uu}	0,590	0,090	0,610	0,110	0,617	0,117

Fonte: Elaborada pelos autores

As simulações evidenciaram que as estimativas do vetor de parâmetros$\theta = (\beta_0, \beta_1, \mu_x, \sigma_{xx}, \sigma_{uu})^T$ possuem menor viés quanto maior é o grau de liberdade na distribuição *t*-Student. Por exemplo, para amostra de tamanho 200 e graus de liberdade igual a 10, o valor da estimativa de σ_{xx} foi de 1,872 com viés de 0,372 (Tabela 3), enquanto que para o caso onde consideramos 40 graus de liberdade, a respectiva estimativa é 1,571 com viés 0,020 (Tabela 6). Esta característica se deve à propriedade da função de distribuição da *t*-Student se aproximar da distribuição Normal quando cresce o grau de liberdade.

Tabela 4 – Estimativas dos parâmetros e respectivos vieses quando usamos a distribuição *t*-Student para gerar Z_i, com $\nu = 20$ graus de liberdade.

Parâmetros	*n* = 20		*n* = 30		*n* = 40	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,351	−0,149	0,425	−0,075	0,445	−0,055
β_1	1,030	0,030	1,015	0,015	1,011	0,011
μ_x	5,001	0,001	4,999	−0,001	5,001	0,001
σ_{xx}	1,622	0,122	1,639	0,139	1,647	0,147
σ_{uu}	0,489	−0,011	0,512	0,012	0,524	0,024
Parâmetros	*n* = 50		*n* = 100		*n* = 200	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,444	−0,056	0,485	−0,015	0,494	−0,006
β_1	1,011	0,011	1,003	0,003	1,001	0,001
μ_x	5,002	0,002	4,999	−0,001	5,000	0,000
σ_{xx}	1,644	0,144	1,661	0,161	1,661	0,161
σ_{uu}	0,531	0,031	0,543	0,043	0,548	0,048

Fonte: Elaborada pelos autores

Em geral, concluimos que, o desempenho das estimativas dos parâmetros do modelo de regressão com erros nas variáveis, é melhor quando a distribuição dos dados segue uma distribuição Normal em comparação com distribuição *t*-Student, como pressuposto pelo modelo. É fundamental condicionar a aplicação do modelo normal com erros nas variáveis quando os dados provêm da *t*-Student, não só ao tamanho amostral suficientemente grande, em nosso caso acima de 50 observações, mas também para os casos com maior grau de liberdade, isto é, maior do que 20, conforme verificado nas simulações.

Tabela 5 – Estimativas dos parâmetros e respectivos vieses quando usamos a distribuição *t*-Student para gerar Z_i, com $\nu = 30$ graus de liberdade.

Parâmetros	n = 20		n = 30		n = 40	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,364	−0,136	0,417	−0,083	0,447	−0,053
β_1	1,028	0,028	1,016	0,016	1,011	0,011
μ_x	5,002	0,002	5,000	0,000	4,998	−0,002
σ_{xx}	1,552	0,052	1,581	0,081	1,583	0,083
σ_{uu}	0,477	−0,023	0,494	−0,006	0,506	0,006
Parâmetros	n = 50		n = 100		n = 200	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,463	−0,037	0,474	−0,026	0,490	−0,010
β_1	1,007	0,007	1,005	0,005	1,002	0,002
μ_x	5,000	0,000	5,001	0,001	4,999	−0,001
σ_{xx}	1,597	0,097	1,596	0,096	1,596	0,096
σ_{uu}	0,511	0,011	0,524	0,024	0,530	0,030

Fonte: Elaborada pelos autores

ANÁLISE DE DADOS BRASILEIROS DA TAXA DE DESEMPREGO NO ANO DE 2019 EM FUNÇÃO DO ÍNDICE DE GINI

Os dados analisados neste trabalho são referentes à taxa de desemprego em função do Índice de Gini, no ano de 2019, nos 26 estados brasileiros. O estudo buscou ajustar a regressão linear simples e a regressão com erros nas variáveis para os dados e, em seguida, selecionar o melhor modelo.

O Índice de Gini é uma medida estatística que avalia o grau de concentração de renda (IPEA, 2004). O cálculo deste índice baseia-se na Curva de Lorenz, que apresenta a relação entre a porcentagem acumulada da população, em ordem crescente de renda, e a porcentagem acumulada de rendimentos. Essa curva é hipotética, que varia de zero à um, com ângulo de 45 graus e representa a igualdade perfeita de renda entre os indivíduos, de forma que quanto mais próximo de zero, menor é a concentração de renda e, na

situação oposta, quanto mais próximo de um, maior é o acúmulo de rendimentos por poucos indivíduos (NISHI, 2010).

Tabela 6 – Estimativas dos parâmetros e respectivos vieses quando usamos a distribuição *t*-Student para gerar Z_i, com $\nu = 40$ graus de liberdade.

Parâmetros	n = 20		n = 30		n = 40	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,373	−0,127	0,409	−0,091	0,435	−0,065
β_1	1,025	0,025	1,018	0,018	1,013	0,013
μ_x	4,996	−0,004	4,998	−0,002	5,004	0,004
σ_{xx}	1,528	0,028	1,553	0,053	1,557	0,057
σ_{uu}	0,467	−0,033	0,489	−0,011	0,500	0,000
Parâmetros	n = 50		n = 100		n = 200	
	Estimativa	Viés	Estimativa	Viés	Estimativa	Viés
β_0	0,454	−0,046	0,486	−0,014	0,490	−0,010
β_1	1,009	0,009	1,003	0,003	1,002	0,002
μ_x	5,000	0,000	5,002	0,002	4,999	−0,001
σ_{xx}	1,561	0,061	1,573	0,073	1,571	0,071
σ_{uu}	0,502	0,002	0,514	0,014	0,520	0,020

Fonte: Elaborada pelos autores

Para obter o Índice de Gini são necessárias informações do tamanho da população e da renda per capita. Neste aspecto que se encontram as possibilidades de erros de mensuração. No caso da população brasileira, o seu tamanho e rendimentos individuais são determinados a cada dez anos por meio do censo populacional ou, ainda, acompanhada de forma contínua na Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua. Os dados de 2019, que utilizamos em nosso trabalho, foram obtidos da PNAD Contínua (IBGE, 2021a) e, de acordo com o que relatamos, podem apresentar erros relacionados à declaração do participante da sua renda e da estimação a partir de uma amostra populacional.

Outro aspecto importante é a variável desemprego, a qual buscamos verificar a sua associação com o Índice de Gini. A taxa de desemprego, também chamada de taxa de desocupação, é obtida pela relação entre às pessoas com idade para trabalhar (maiores de 14 anos de idade), mas que não estão trabalhando embora estejam em busca de emprego. Logo, universitários que se dedicam exclusivamente aos estudos e uma dona de casa que não trabalha fora, por exemplo, não são considerados indivíduos desocupados (IBGE, 2021b).

Assim como o Índice Gini, as informações da taxa média de desemprego de 2019 dos estados brasileiros foram retiradas dos "Estudos especiais do Banco Central: O desalento e as taxas de desocupação", que se fundamentaram em dados da PNAD Contínua (BANCO CENTRAL, 2020). Este fator, nos leva a compreender que a base de dados pode estar sujeita a erros referentes a coleta de informações, pois o indivíduo pode omitir ou supervalorizar a sua situação empregatícia.

Diante do exposto, aqui buscamos analisar a regressão com erros nas variáveis e compará-la com a regressão linear simples em que o Índice de Gini é a variável explicativa e a taxa de desocupação, a variável resposta.

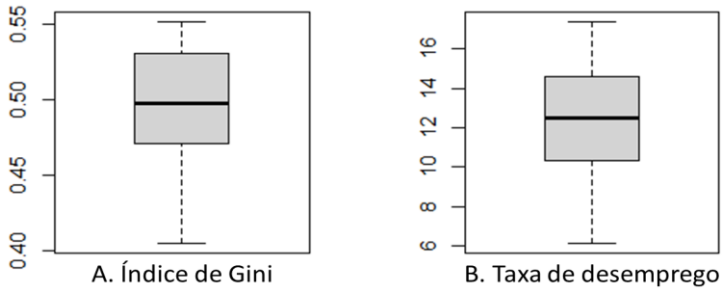
Na estimação dos parâmetros nos modelos de regressão clássico foi utilizado o método de mínimos quadrados e no modelo com erros de medida, as estimativas foram obtidas pelo método dos momentos, dadas na Equação (7).

Para a análise da regressão linear simples, verificou-se os pressupostos de distribuição Normal dos resíduos pelo Teste de Shapiro-Wilk (RAZALI; WAH, 2011, MIOT, 2017). Para a regressão com erros nas variáveis, analisou-se a presença de distribuição Normal bivariada pelo Teste de Royston (ROYSTON, 1983, CANTELMO; FERREIRA, 2007). A seleção do modelo com melhor ajuste foi realizada a partir dos critérios de seleção de Akaike (1973) (AIC) e Bayesiano (BIC) de Schwartz (1978). O modelo escolhido será aquele com menor valor de AIC e BIC, sendo que, para valores próximos, a seleção será fundamentada nos aspectos relatados de parcimônia, consistência da técnica utilizada e sua eficiência em relação aos resultados obtidos.

Nos dados utilizados, temos que a média do Índice de Gini é 0,50 e a taxa média de desemprego é de 12,10%. Além disso, temos que os menores valores para o Índice de Gini e taxa de desemprego são, respectivamente, 0,41 e 6,10% e os maiores são 0,55 e 17,40%.

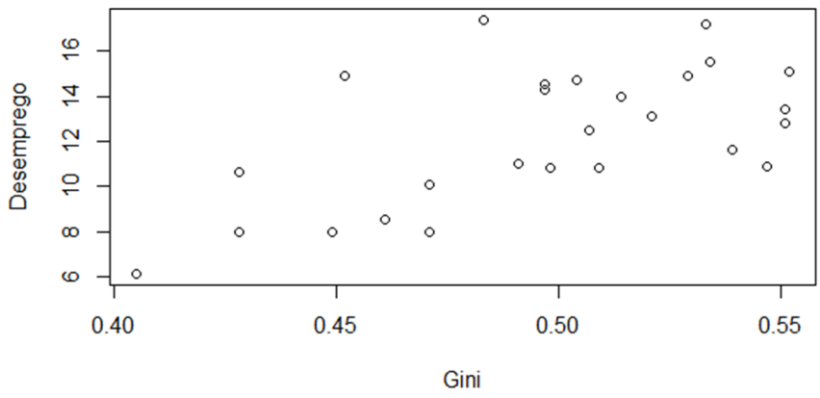
Não foram observados pontos extremos (Figura 1) e o teste de Shapiro-Wilk não rejeitou a normalidade das variáveis Índice de Gini e taxa de desemprego, já que os valores-*p* são 0,27 e 0,35, respectivamente. O teste de Royston evidenciou que a distribuição conjunta dessas variáveis é Normal bivariada. A Figura 2 mostra a dispersão dos dados. Observamos que à medida que aumentam os valores de Índice de Gini também aumenta a taxa de desemprego. Tal associação pode ser evidenciada pela análise de correlação de Pearson (r = 0,61, valor-p <0.001) (SCHOBER et al, 2018).

Figura 1 – Boxplot dos dados de Índice de Gini (A) e taxa de desemprego (B) em 2019.



Fonte: Elaborada pelos autores

Figura 2 – Dispersão da taxa de desemprego em relação ao Índice de Gini em 2019.



Fonte: Elaborada pelos autores

Na Tabela 7, apresentamos as estimativas dos parâmetros nos casos considerados, a saber: modelos com e sem erros de medida. Observamos que o coeficiente β_1 foi significativo em ambos modelos (valor-p < 0,01). Temos uma modificação importante nas estimativas dos parâmetros da regressão, veja por exemplo, que $\hat{\beta}_1$ é igual a 45,39 (modelo sem erros nas variáveis) e 122,35 (modelo com erros nas variáveis). Na comparação dos modelos pelos critérios de seleção, encontramos que a regressão com erros nas variáveis apresenta melhor ajuste segundo os critérios AIC e BIC em relação ao modelo sem erros nas variáveis. Por exemplo, para o modelo com erros nas variáveis o valor de AIC obtido foi de –118,52 enquanto para a regressão simples foi de 129,47.

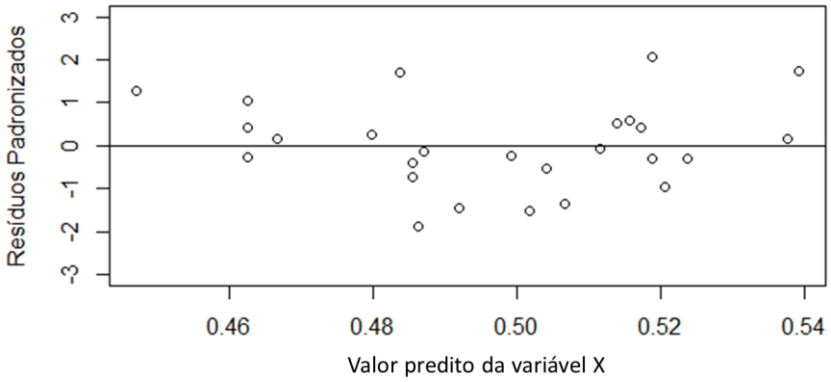
Tabela 7 – Estimativas dos parametros e critérios de seleção referentes à regressão linear sem e com erros.

	Modelo sem erros nas variáveis	Modelo com erros nas variáveis
Intercepto $(\hat{\beta}_0)$	-10,418	-48,590
Índice de Gini $(\hat{\beta}_1)$	45,395	122,349
Critério de seleção AIC	129,472	-118,519
Critério de seleção BIC	133,360	-115,928

Fonte: Elaborada pelos autores

Uma vez selecionado o modelo mais adequado para os dados, realizamos o diagnóstico da qualidade do ajuste. Visualmente, a Figura 3 nos mostra que os resíduos padronizados não apresentam tendência com os valores preditos da variável explicativa ou mesmo variação na sua dispersão.

Figura 3 – Análise dos resíduos padronizados da regressão com erro nas variáveis.



Fonte: Elaborada pelos autores

CONSIDERAÇÕES FINAIS

Este trabalho insere-se no contexto de informações que são observadas de forma indireta e, por isso, podem apresentar erros de medida. Ignorar esses erros pode gerar modelos de regressão viesados e com baixa qualidade na explicação do comportamento dos dados. Desta forma, estabelecemos o objetivo de analizar os estimadores do modelo normal com erros nas variáveis quando os dados provêm de uma população com distribuição Normal ou *t*-Student. A partir dos estudos de siulação, evidenciamos que as estimativas do modelo normal com erros nas variáveis são menos viesadas quando o pressuposto de normalidade da distribuição está presente e que, quando violamos esse pressuposto, ou seja, usamos a distribuição *t*-Student para gerar os dados, os vieses diminuem à medida que aumentamos o tamanho amostral e/ou os graus de liberdade, corroborando a literatura existente.

Além disso, apresentamos uma aplicação a dados reais onde ajustamos um modelo com erros nas variáveis e outro sem erros nas variáveis, e verificamos aquele com melhor ajuste segundo os critérios AIC e BIC. As variáveis consideradas foram Índice de Gini e taxa de desemprego brasileiro no ano de 2019. Neste caso, o modelo com erros nas variáveis apresentou melhor ajuste, usando os critérios de seleção. Assim, a aplicação corrobora a hipótese levantada de que este método é mais adequado para a análise de informações passíveis de erros. Finalmente, sugerimos que o modelo com erro nas variáveis deva ser considerado para situações semelhantes, em que as variáveis independentes podem conter erros associados a elas.

REFERÊNCIAS

AKAIKE, H. **Information Theory and an Extension of the Maximum Likelihood Principle**. In B. N. Petrov, & F. Csaki (Eds). Em Second International Symposium on Information Theory. Akademiai Kiado, Budapest, p. 276-281, 1973.

AL-SHARADQAH, A.; CHERNOV, N.; HUANG, Q. **Errors-In-Variables regression and the problem of moments**. Brazilian Journal of Probability and Statistics, v. 27, n. 4, 2013.

BABA, M. Y. **Inferência bayesiana para regressão linear simples com erro nas variáveis**. Mestrado em Ciências de Computação e Matemática Computacional—São Carlos: Universidade de São Paulo, 4 jul. 2018.

BANCO CENTRAL. **O desalento e as taxas de desocupação. Estudos Especiais do Banco Central,** n.78, p. 1-3, abr. 2020. Disponível em: <https://www.bcb.gov.br/conteudo/relatorioinflacao/EstudosEspeciais/EE078_O_desalento_e_as_taxas_de_desocupacao.pdf>. Acesso em: 21 de agosto de 2022.

BICKEL, P. J.; RITOV, Y. **Efficient Estimation in the Errors in Variables Model**. The Annals of Statistics, v. 15, n. 2, 1987.

CANTELMO, N. F.; FERREIRA, D. F. **Desempenho de testes de normalidade multivariados avaliado por simulação Monte Carlo**. Ciência e Agrotecnologia, v. 31, n. 6, p. 1630–1636, 2007.

CARRASCO, J. M. F. **Modelos de regressão beta com erro nas variáveis**. Doutorado em Estatística - São Paulo: Universidade de São Paulo, 2012.

CARVALHO JÚNIOR, J. G.; ALMEIDA, S. S.; RAMOS, E. M. L. S. **Gráfico de Controle de Regressão Estrutural**. TEMA - Tendências em Matemática Aplicada e Computacional, v. 8, n. 3, p. 361–370, 2007.

CHENG, C.-L.; VAN NESS, J. W. **On the unreplicated ultrastructural model**. Biometrika, v. 78, n. 2, p. 442–445, 1991.

DE OLIVEIRA, T. A.; DE MORAES, A. R.; CIRILLO, M. A. **Comparação de métodos de estimação em um modelo linear simples com erro nas variáveis**. Ciência e Natura, v. 32, n. 2, p. 23-34, 2010.

DOLBY, G. R. **The ultrastructural relation: A synthesis of the functional and structural relations**. Biometrika, v. 63, n. 1, p. 39-59, 1976.

FULLER, W. A. **Measurement error models**. New York: Wiley, p. 1-99, 1987.

GILLARD, J. **An overview of linear structural models in errors in variables regression**. Statistical Journal, v. 8, n. 1, p. 57–80, 2010.

IBGE. **PNAD Contínua – Pesquisa Nacional por Amostra de Domicílios Contínua**. Instituto Brasileiro de Geografia e Estatística, 2021a <https://www.ibge.gov.br/estatisticas/multidominio/condicoes-de-vida-desigualdade-e-pobreza/17270-pnad-continua.html?=&t=downloads>. Acesso em: 21 de agosto de 2022.

IBGE. **Desemprego**. Instituto Brasileiro de Geografia e Estatística, 2021b <https://www.ibge.gov.br/explica/desemprego.php>. Acesso em: 21 de agosto de 2022.

IPEA. **O que é Índice de Gini?**. Instituto de Pesquisa Econômica Aplicada, 2004 <https://www.ipea.gov.br/desafios/index.php?option=com_content&id=2048:catid=28> . Acesso em: 21 de agosto de 2022.

LEAMER, E. E. **Errors in Variables in Linear Systems**. Econometrica, v. 55, n. 4, 1987.

MIOT, H. A. **Avaliação da normalidade dos dados em estudos clínicos e experimentais**. Jornal Vascular Brasileiro, v. 16, n. 2, p. 88–91, 2017.

NGHIEM, L.; BYRD, M.; POTGIETER, C. **Estimation in linear errors-in-variables models with unknown error distribution**. Biometrika, v. 107, n. 4, p.841-846, 2020.

NISHI, L. F. **Coeficiente de Gini: uma medidade de distribuição de renda**. Universidade do Estado de Santa Catarina, 2010.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2023.

RAZALI, N. M.; WAH, Y. B. **Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests**. Journal of Statistical Modeling and Analytics, p. 1-13, 2011.

REILLY, P. M.; PATINO-LEAL, H. **A Bayesian Study of the Error-in-Variables Model**. Technometrics, v. 23, n. 3, 1981.

ROYSTON, J. P. **Some Techniques for Assessing Multivarate Normality Based on the Shapiro- Wilk W**. Applied Statistics, v. 32, n. 2, 1983.

SCHWARZ, G. **Estimating the dimension of a model**. The Annals of Statistics, v. 6, p. 461–464, 1978.

SOARES, C. P. M. **Uma Abordagem via Mistura Finita para Modelos de Regressão Linear com Erro nas Variáveis**. Mestrado em Estatística, Belo Horizonte: Universidade Federal de Minas Gerais, 2020.

SCHOBER, P.; BOER, C.; SCHWARTE, L. A. **Correlation Coefficients: Appropriate Use and Interpretation**. Anesthesia & Analgesia, v. 126, n. 5, p. 1763–1768, 2018.

TOMAYA, L. Y. C. **Inferência em modelos de regressão com erros de medição sob enfoque estrutural para observações replicadas**. Mestrado em Estatística, São Carlos: Universidade Federal de São Carlos, p. 92, 2014.