Comparação de métodos de estimação em um modelo linear simples com erro nas variáveis

Tiago Almeida de Oliveira[1], Augusto Ramalho de Moraes[2], Marcelo Angelo Cirillo[2]

[1]Doutorado em Estatística e Experimentação Agronômica
ESALQ/USP - Piracicaba, SP
[2]Departamento de Ciências Exatas
Universidade Federal de Lavras, Lavras, MG
e-mail: macufla@gmail.com

Resumo

Por mais cauteloso que o pesquisador seja, os erros nas variáveis poderão estar presentes em um experimento, portanto, é coerente avaliar a confiabilidade das medidas, principalmente no que tange ao uso de diferentes métodos de estimação. Considerando esse fato, este trabalho objetivou comparar o desempenho de dois métodos de estimação, denominados por "Plug-in" e Atenuador de Vício, utilizados para estimar parâmetros de um modelo de regressão com erro nas variáveis. Para isso, foi utilizada a técnica de simulação Monte-Carlo, da qual foi gerado um modelo de regressão linear simples com erro nas variáveis, submetido a diferentes situações representadas pelos tamanhos amostrais, distribuição dos resíduos e valores referentes à qualidade de ajuste. Concluiu-se que, em alta qualidade de ajuste (R^2=90%), os métodos "Plug-in" e Atenuador de Vício foram precisos, porém acurados apenas para os modelos com distribuição simétrica.
Palavras-chave: erros de medida, regressão linear, viés.

Abstract

Although the researcher is very careful errors experiment may be present, therefore, is appropriate to evaluate the reliability of data, mainly when different estimation methods are used. Considering this problem, the objective of this work is compare the performance of two methods of estimation known as "Plug-in" and "Attenuation bias". These methods are used in the regression model with errors-in-variables techniques. The Monte-Carlo simulations are used for generated a linear regression model errors-in-variables, different sample sizes, residuals distributions and lack of fit were considered. It was possible to conclude that, "Plug-in " and

"Attenuation bias" methods were accurate with R^2=90%, however, these methods have better accuracy for symmetrical distributions.
Keywords : errors-in-variables, linear regression, bias

Introdução

Em geral, os modelos de regressão são ajustados sobre a hipótese de que a covariável é medida sem erro, ou seja, considerada como efeito fixo. Porém, em algumas situações, a amostragem feita com repetição poderá revelar que essa covariável venha apresentar efeitos aleatórios, uma vez que fatores do acaso poderão interferir no processo de medição. Uma situação prática a ser exemplificada nesse contexto refere-se à medição de uma massa com uma balança. Nesse caso, fatores aleatórios, como correntes de ar ou vibrações, introduziriam erros nas mensurações. Uma alternativa para evitar essa alteração seria repetir inúmeras vezes a medição, permitindo avaliar a incerteza estatística no resultado final.

Situações desse tipo se enquadram na engenharia, na qual a redução dos "erros", no contexto mais amplo da palavra, é tratada especificamente pela ciência denominada por metrologia. Em síntese, essa ciência analisa os aspectos relativos às medições e calibrações de instrumentos, utilizando-se metodologias estatísticas. Tais medições, em conjunto com o uso de uma técnica estatística inadequada certamente resultam em índices de qualidade incoerentes.

Em virtude de que a medição de um resultado está sujeito a pequenas variações aleatórias, nota-se que o enfoque estatístico consistirá na modelagem de dados, considerando o erro aleatório nas variáveis independentes (Vuolo, 1992). Dessa forma, um modelo que contemple o erro de medida agregará informações mais precisas dentro da confiabilidade metrológica, quando comparada à modelagem feita de forma convencional, assumindo o efeito fixo nas covariáveis.

Na presença do erro de medição, o valor verdadeiro da variável dependente é considerado como um valor não observado, pois supostamente assume-se que ele encontra-se contaminado por algum erro. Contudo, nem sempre é fácil determinar a distribuição estatística desse erro e, por isso, torna-se complexa a inferência, bem como a formulação de novos métodos de estimação. Uma alternativa para solucionar esse problema seria o uso de técnicas de simulação Monte Carlo, aliados aos métodos de computação intensiva.

Alguns trabalhos com importantes contribuições, proporcionadas por essas técnicas, são os de Higdon e Schafer (2001) em que os autores apresentam métodos computacionais baseados em algoritmos E-M (Dempster et al. 1977) e a modificação no método da quadratura Gauss-

Hermite (Liu e Pierce, 1994), utilizados na obtenção dos estimadores de máxima verossimilhança, nos modelos lineares generalizados com erro de medida nas covariáveis. Com o propósito de comparar o desempenho desses métodos, os autores observaram inacurácia na quadratura gaussiana e problemas de singularidade nesse procedimento numérico.

Em relação à qualidade de ajuste de modelos com erros de medida, Delicato e Romo (2004) propuseram testes estatísticos paramétricos, apresentando suas distribuições assintóticas e intervalos de confiança para os parâmetros, utilizando-se a técnica *bootstrap*. Esses testes foram validados por meio das probabilidades empíricas, baseando-se no controle do erro tipo I e poder.

Ao se tratar dos métodos de estimação dos parâmetros de um modelo com erro de medida, uma particular atenção é dada para os estimadores baseados em métodos dos momentos, desenvolvidos por Fuller (1987) e por Stein e James (1961), sendo utilizados para estimar o verdadeiro valor da variável independente. Respectivamente, os métodos citados são conhecidos na literatura como estimadores "Plug-in" e de Atenuação de Vício. Ambos os estimadores foram desenvolvidos com a finalidade de minimizar os riscos de incerteza, minimizando o efeito do erro de medida.

Basicamente, a diferença na formalização desses métodos, citada por Cunha e Colosimo (2003), é que os estimadores "Plug-in" estimam o valor verdadeiro através dos valores observados e, de posse dos respectivos valores ajustados, utilizam a função de estimação usual no modelo estudado para estimar o parâmetro de interesse. Já no caso dos estimadores conhecidos como Atenuador de Vício, sabe-se que eles apenas realizam uma correção no vício a partir do estimador de mínimos quadrados clássico.

A construção de intervalos de confiança para os parâmetros, considerando esses métodos, depara-se com o problema de obter suas variâncias assintóticas, uma vez que o processo de estimação é realizado em duas etapas, conforme é relatado por Cunha e Colosimo (2003), que propuseram intervalos de confiança pela abordagem *bootstrap*. Destacando-se o intervalo obtido pelo método do percentil (Davison e Hinkley, 1997).

Com o propósito de comparar os métodos de estimação para modelos de regressão com erro de medida na variável, objetivou-se avaliar os vieses proporcionados pelos estimadores "Plug-in" e Atenuador de Vício, realizando um estudo de Monte Carlo em um modelo de regressão linear simples.

Metodologia

A metodologia proposta neste trabalho considerou um modelo de regressão com erro nas variáveis, o modelo de análise é semelhante ao

mencionado por Searle (1971). Salienta-se que a expansão dos métodos de estimação para modelos mais complexos do ponto de vista estatístico é perfeitamente cabível dentro da metodologia exposta, segue-se o modelo.

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \ (i = 1,...,h) \text{ e } (j = 1,...k), \text{em que} \qquad (1)$$

y_{ij} correspondeu aos valores da variável dependente na *i-ésima* unidade amostral e *j-ésima* repetição; ε_{ij} o erro aleatório, simulado nesse trabalho sob as distribuições: Normal ($\varepsilon_{ij} \sim U(0,1)$); Uniforme ($\varepsilon_{ij} \sim U(0,1)$) e Gama ($\varepsilon_{ij} \sim Ga(4,1)$).

A variável independente foi representada pelo vetor $X^t=(x_{11},...,x_{bk})^t$, em que, *k* correspondeu ao número total de repetições, assumindo, os valores *4; 6; 10; 14* e *20*; *h* correspondeu ao número total de unidades amostrais, previamente fixado em $h=5$. Assim sendo, o tamanho da amostra foi determinado por *hxk*, sendo constituído por $n = 20; 30; 50; 70$ e *100* unidades amostrais.

Cada componente de (x_{ij}) $(i = 1,...,h)$ e (j=1,...,k) representou um valor fixo observado, desconsiderando o erro de medida. Nesse caso, a estimação dos parâmetros do modelo (1) foi realizada por meio do método de mínimos quadrados.

O pesquisador pode ter motivos para acreditar que o valor observado da covariável não corresponde ao valor exato, contendo algum erro, que pode ser devido à calibração do aparelho escolhido para fazer as aferições (Souza, 2004). O erro de medida pode ser incorporado em um modelo linear, por meio da relação:

$$z_{ij} = x_{ij} + u_{ij} \text{ , em que} \qquad (2)$$

z_{ij} representou a covariável, medida com erro na *i-ésima* unidade amostral e *j-ésima* repetição, gerada através de uma distribuição uniforme (0,1). Em relação ao valor do erro de medida (u_{ij}), ele foi calculado pela diferença, uma vez que x_{ij} referiu-se a um valor já conhecido. A qualidade de ajuste do modelo foi garantida no processo de simulação, assumindo diferentes valores de R^2 previamente fixados em *20, 50, 70* e *90%* sob a restrição de que a soma de quadrados residual (SQE) foi determinada sob a igualdade.

$$SQE = \frac{\left[\left(1 - R^2\right) \times SQR\right]}{R^2} \qquad (3)$$

sendo que SQR correspondeu à soma de quadrados de regressão.

Para cada situação configurada por diferentes tamanhos amostrais,

distribuições dos resíduos e valores de R^2, as estimativas dos parâmetros do modelo (1) foram obtidas pelos métodos de estimação "Plug-in" e Atenuador de Vício, nos quais, foi possível calcular os vieses referentes às estimativas dos parâmetros, onde, arbitrariamente o modelo simulado foi dado com os valores paramétricos $\beta_0 = 1$ e $\beta_1 = 1$. Em relação ao estudo da predição, obteve-se o viés médio (Vm), permitindo interpretar a acurácia média das predições:

$$Vm = \frac{\sum_{l=1}^{N} (\hat{y}_l - y_l)}{N}, \text{ sendo} \qquad (4)$$

N representou o número total de simulações realizadas.

Estimador "Plug-in" proposto por Stein e James

Supondo que os valores z_{ij} *(i=1,...h; j=1,...,k)* da covariável, observada com o erro de medida, foram obtidos nas *k* repetições, sendo essas, independentes e identicamente distribuídas por uma distribuição normal com média x_i *(i=1,...,5)* e variância σ_u^2. O estimador proposto por Stein e James (1961) parte do pressuposto em ajustar um novo valor para a covariável z_{ij}. Para isso, utilizou-se a expressão:

$$e_i(z) = \hat{B}\bar{z} + (1 - \hat{B})\bar{z}_{i.} \quad \text{com } 1=1, \ldots,h, \text{ sendo que:} \qquad (5)$$

$\bar{z}_{i.}$ correspondeu à média das k repetições (6); S_u^2 referiu-se ao estimador da variância do erro de medida (7) e $\hat{S}$ à estimativa da distância do valor observado para a média das repetições $\bar{z}$ (9).

$$\bar{z}_{i.} = \frac{\sum_{j=1}^{k} z_{ij}}{k} \qquad (6) \qquad S_u^2 = \frac{\sum_{i=1}^{5} (z_{ij} - \bar{z}_{i.})^2}{hk(k-1)} \qquad (7)$$

$$\hat{B} = \frac{\hat{S}_u (n-3)}{\hat{S}} \qquad (8) \qquad \hat{S} = \sum_{i=1}^{5} (z_{ij} - \bar{z}_{i.})^2 \qquad (9)$$

$$\text{e} \quad \bar{z} = \frac{\sum_{i=1}^{5} \sum_{j=1}^{k} z_{ij}}{hk}$$

Estimador Atenuador de Vício proposto por Fuller

O estimador proposto por Fuller (1987) baseou-se na determinação do vício produzido pelo erro na observação dos dados. O Atenuador do vício foi determinado inicialmente calculando o fator.

$$k_x^{-1} = \frac{\sigma_x^2 + \sigma_u^2}{\sigma_x^2} \qquad (10)$$

podendo ser estimado por

$$k_x^{-1} = \frac{S_z^2}{S_z^2 - S_u^2} \qquad (11)$$

em que S_z^2 correspondeu a variância amostral da covariável ajustada com o erro de medida.

A estimativa de menor vício para cada parâmetro do modelo foi obtida calculando

$$\hat{\beta}_F = \hat{\beta} \times k_x^{-1} \qquad (12)$$

$\hat{\beta}$ é a estimativa de mínimos quadrados.

Com base no exposto acima, a finalização dessa metodologia se deu com a construção de um programa no software R, versão 2.10.1, onde foi implementado o método Monte Carlo. Realizou-se 5000 simulações para obtenção dos resultados.

Resultados e discussão

Uma questão atrativa ao estudo de vieses de estimadores de regressão refere-se ao fato de interpretar o viés como um quantificador do desconhecimento técnico do pesquisador a respeito do instrumento utilizado na realização do experimento. Nesse contexto, Pereira (2004) comenta que a confiabilidade das medidas, seja por instrumentos não calibrados, seja por outra causa de variação controlável, pode vir a ser avaliada por meio do viés, assim sendo, a diferenciação dos resultados, provenientes de diferentes métodos de estimação, pode eventualmente levantar dúvidas sobre essa interpretação, se o estimador com menor viés não for utilizado.

Tendo em vista essa importância e seguindo as especificações descritas na metodologia, inicialmente analisaram-se os vieses médios para os métodos em questão. Para isso, os resultados foram obtidos considerando que a variável dependente foi gerada de tal forma que o modelo explicasse diferentes porcentagens da variabilidade total, quantificadas pelos valores

do coeficiente de determinação R^2 (*20%* e *90%*).

Convém salientar que a acurácia e precisão, de certa forma, encontram-se relacionadas, no sentido de que, dado um valor elevado do viés médio, como consequência, ocasionará um aumento na estimativa do erro residual, levando a um modelo que apresente uma deficiência na precisão. Nas Figuras 1-4, estão os vieses médios para os métodos de estimação estudados.

Com o propósito de avaliar o desempenho do estimador "Plug-in" em um modelo que representasse uma situação de baixa qualidade de ajuste (R^2=20%), gerou-se o modelo linear (Figura 1). Comparando-se o viés médio obtido no ajuste dos três modelos diferenciados pela distribuição dos resíduos: Normal (0,1); Uniforme (0,1) e Gama (4,1), observou-se que, no caso das amostras menores (n=20), os modelos foram menos acurados, pois apresentaram viéses superiores a 0,10. Notoriamente, essa baixa acurácia foi mais pronunciada ao considerar o modelo com os resíduos dados pela distribuição Gama (4,1). Aumentando o tamanho amostral, observou-se uma melhoria na acurácia em virtude de que os resultados evidenciaram uma redução próxima a 0,05.

Aumentando o valor da qualidade de ajuste (R^2=90%), os resultados (Figura 2) confirmaram que os modelos com os resíduos distribuídos por uma Normal (0,1) e Uniforme (0,1) foram mais acurados, incluindo até mesmo as amostras menores. Esse fato não foi verificado no modelo com os resíduos distribuídos por uma Gama (4,1), portanto, há evidências para afirmar que o modelo de regressão com erros de medida, cujas estimativas foram obtidas por meio do estimador "Plug-in" é mais acurado nas situações com elevada explicação da variabilidade dos dados amostrais, preferivelmente, com os resíduos dados pelas distribuições simétricas.

Tratando-se das estimativas dos parâmetros dos modelos, obtidas pelos estimadores de Fuller, denominados como Atenuadores de Vício, foi verificado que, no caso do modelo de regressão gerado com baixa qualidade de ajuste, R^2=20% (Figura 3), o viés médio apresentou um comportamento semelhante ao obtido pelo estimador "Plug-in" (Figura 1), porém com valores inferiores. Especificamente, no modelo de regressão com resíduos normais, a diferença nesses valores, quando comparados ambos os métodos, aproxima-se de 0,05. Embora possa parecer um resultado desprezível, levar em consideração alguns aspectos inferenciais, tais como testes de hipóteses que envolvam estimativas dos parâmetros, supostamente, a confiabilidade dos resultados poderá ser comprometida, de acordo com o método de estimação utilizado.

Dada uma alta qualidade de ajuste R^2 =90% (Figura 4), foi notado que os valores dos vieses médios, para todos os tamanhos amostrais avaliados, foram similares aos resultados proporcionados pelo estimador "Plug-in" (Figura 2).

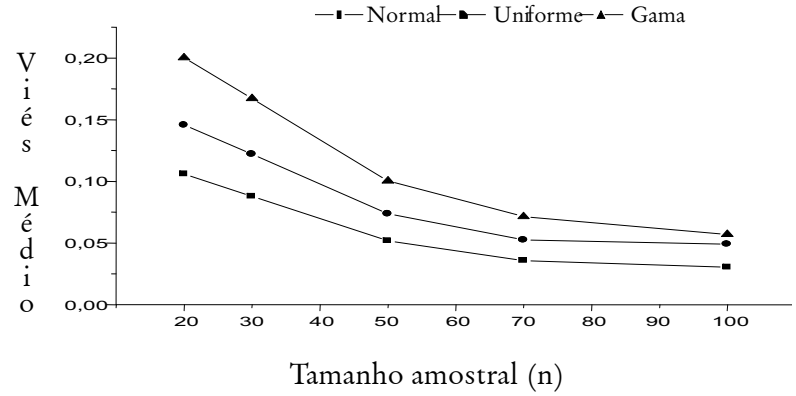

Figura 1. Viés médio do estimador "Plug-in" considerando o modelo linear com os resíduos distribuídos por: Normal(0,1), Uniforme(0,1), Gama (4,1) e coeficiente de determinação R^2=20%.

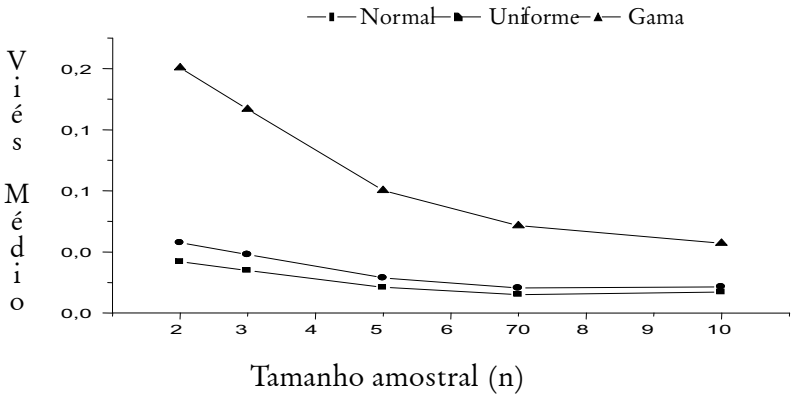

Figura 2. Viés médio do estimador "Plug-in" consideran-do o modelo linear com os resíduos distribuídos por: Normal (0,1),Uniforme(0,1) e Gama(4,1) e coeficiente de determinação R^2=90%.

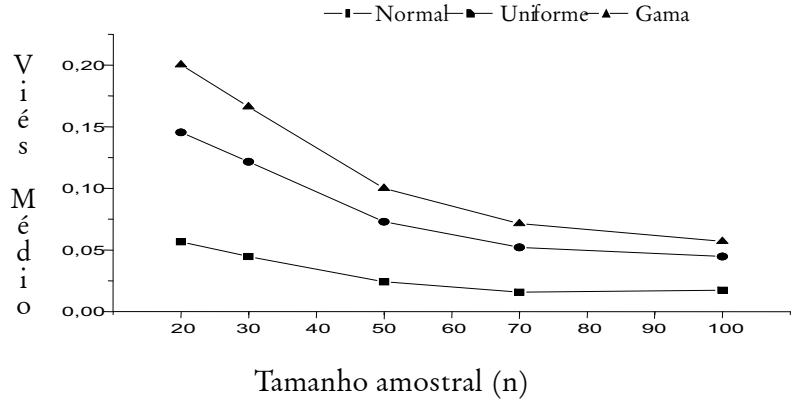

Figura 3. Viés médio do estimador Atenuador de Vício, considerando o modelo linear com os resíduos distribuídos por: Normal (0,1); Uniforme (0,1) e Gama(4,1) e coeficiente de determinação R^2=20%.

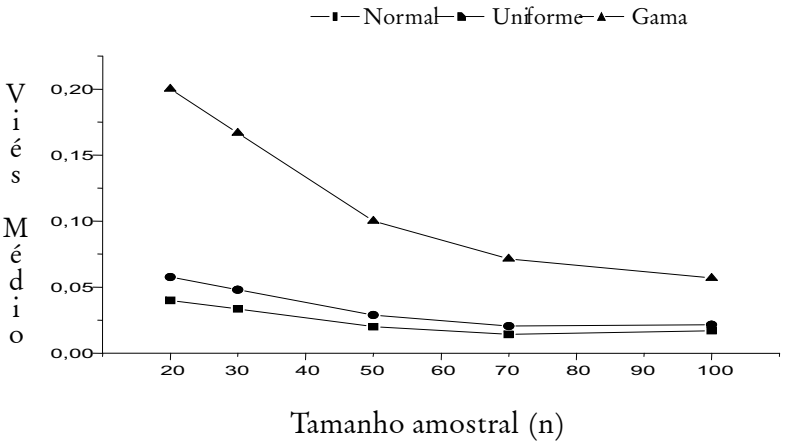

Figura 4. Viés médio do estimador Atenuador de Vício, considerando o modelo linear com os resíduos distribuídos por: Normal (0,1); Uniforme (0,1) e Gama(4,1) e coeficiente de determinação R^2=90%.

Pelos resultados anteriores, constatou-se que os modelos avaliados foram mais acurados para grandes amostras. Assim sendo, em particular, consideraram-se as amostras maiores (n=100) para estudar os vieses das estimativas dos parâmetros para os dois estimadores: "Plug-in" e Atenuador de Vício (Tabela 1).

Tabela 1. Viés das estimativas dos parâmetros do modelo linear sob diferentes valores de R^2, fixando o tamanho amostral (n=100), para diferentes distribuições dos erros nos dois métodos de estimação.

R^2 (%)	M.E.*	$\varepsilon \sim N(0,1)$		$\varepsilon \sim U(0,1)$		$\varepsilon \sim Ga(4,1)$	
		β_0	β_1	β_0	β_1	β_0	β_1
90	PL[1]	0,0294	0,0018	0,0919	0,0004	7,8122	0,0127
	AV[2]	0,0008	0,0002	0,0926	0,0000	8,0112	0,0010
70	PL[1]	0,0507	0,0039	1,7921	0,0016	8,1134	0,0002
	AV[2]	0,0003	0,0004	1,8177	0,0000	8,0008	0,0002
50	PL[1]	0,0257	0,0020	2,7674	0,0004	7,9837	0,0019
	AV[2]	0,0181	0,0010	2,7738	0,0000	7,9982	0,0005
20	PL[1]	0,1169	0,0080	5,4812	0,0048	7,9082	0,0059
	AV[2]	0,0224	0,0010	5,5538	0,0002	8,0154	0,0012

Ao avaliar os resultados referentes aos vieses das estimativas do coeficiente de inclinação (β_1) (Tabela 1), verificou-se que ambos os métodos de estimação foram acurados e precisos. Esse fato foi notório para todas as situações envolvendo os modelos nas diferentes distribuições dos resíduos e qualidades de ajuste, simuladas conforme os valores do R^2 previamente fixados.

Em relação aos resultados referentes aos vieses do intercepto (β_0), foram percebidos valores mais discrepantes ao comparar ambos os métodos. Tal discrepância foi mais pronunciada na situação em que o modelo foi simulado com baixa qualidade de ajuste (R^2=20%) e com os resíduos normalmente distribuídos. Nessa situação, observou-se que o estimador "Plug-in" apresentou um viés de 0,1169, enquanto o estimador Atenuador de Vício propiciou um viés no valor de 0,0224, uma diferença de 0,0945. Além do mais, os resultados propiciados pelo estimador Atenuador de Vício foram inferiores. Um resultado peculiar a esse estimador, ainda mantendo-se

o modelo com distribuição do resíduo normal ($\varepsilon_{ij} \sim N(0,1)$), foi que, para os modelos com alta explicação da variabilidade, isto é, $R^2=70$ e 90%, observou-se uma maior precisão, de forma que os valores passaram a se diferenciar na quarta casa decimal. À medida que se reduziu o coeficiente de determinação ($R^2=50$ e 20%) essa diferenciação ocorreu a partir da segunda casa decimal.

Quando o modelo foi simulado com os resíduos distribuídos por uma uniforme U (0,1) e Gama (4,1), observou-se que a precisão proporcionada pelos estimadores "Plug-in" e Atenuador de Vício apresentaram uma pequena oscilação a partir da segunda casa decimal. Esse fato se deu para todos os coeficientes de determinação simulados. Vale ressaltar que esses resultados retrataram uma falta de acurácia dos métodos em questão, mediante essas distribuições.

Conclusões

O método proposto por Fuller apresentou melhor desempenho do que o estimador "Plug-in" sobre distribuição simétrica, tanto na acurácia quanto na precisão, nos diferentes tamanhos amostrais, em situações de baixa qualidade de ajuste ($R^2=20\%$).

A acurácia dos estimadores "Plug-in" foi sensível à distribuição dos resíduos e diferentes valores de R^2 e tamanhos amostrais dos modelos lineares.

O viés da estimativa do parâmetro intercepto (β_0), para os estimadores "Plug-in" e Atenuador de Vício, apresentou valores discrepantes em algumas situações, revelando baixa acurácia. No caso do coeficiente de inclinação, ambos os métodos apresentaram praticamente o mesmo desempenho.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de estudos.

Referências bibliográficas

BIBBY, J.; TOUTENBURG, H. **Prediction and improved estimation in linear models**. John Wiley, London, 1977.

CUNHA, W.J. ; COLOSIMO E. A.Intervalos de confiança *bootstrap* para modelos de regressão com erros de medida. **Revista de Matemática**

e Estatística. São Paulo, v.21, n.2, 2003.

DAVISON, A.C.; HINKLEY, D.V. **Bootstrap methods and their application**. New York: Cambridge University Press, 1997.

DEMPSTER, A.P. ; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). **Journal Royal Statistic Society. Ser. B** 39, 1977.

DELICATO, P.F. ; ROMO, J. Goodness-of-fit tests in random coefficient regression models. **Annals Institue Statistics Mathematic,** 51, 1999.

DELICATO, P.F. ; ROMO, J. Random coefficient regressions: parametric goodness-of-fit-tests. **Journal Statistics Planning Inference**, 119, 2004.

FULLER, W.A. **Measurement Error Models**, Wiley, New York, 1987.

HIGDON, R.; SCHAFER, D.W.Maximum likelihood computations for regression with measurement error, **Computational Statistics Data Analysis**, 35, 2001.

LIU, Q.; PIERCE, D.A. A note on gauss hermite quadrature. **Biometrics**, 81, 1994.

PEREIRA, J.C.R. **Análise de dados qualitativos: Estratégias metodológicas para as ciências da saúde, humanas e sociais**, Edusp, São Paulo, 2004.

R Development Core Team R**:** A language and environment for statistical computing*. R* Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. 2010

SEARLE, S.R. **Linear models**. New York, John Wiley & Sons, 1971. 532p.

SOUZA, F.G. **Análise de covariância com um fator e erro de medida na Covariável.** São Paulo: IME-USP, Junho 2004. Dissertação de Mestrado.

STEIN, C. ; JAMES, W. Estimation with quadratic loss. In: Berkeley Symposium on mathematics, statistics and probability, 4, 1961, Berkeley. **Proceedings**... Berkeley: University of California Press, 1, 1961.

VUOLO, J.H. **Fundamentos da teoria dos erros,** Edgard Blücher, 1992.

Submetido em: 02/setembro/2009
Aceito em: 23/agosto/2010