Regressão Linear com erros-nas-variáveis: Comparação dos estimadores de mínimos quadrados com os estimadores de 3 modelos de tipo II e de máxima verosimilhança

João Maroco
Departamento de Estatística e Unidade de Investigação em Eco-Etologia.
Instituto Superior de Psicologia Aplicada. Rua Jardim do Tabaco, 34. 1149-041 Lisboa.
jpmaroco@ispa.pt

Resumo: Os modelos de regressão linear de tipo I, que consideram erros de medição e/ou variabilidade natural apenas na variável dependente, são frequentemente utilizados em investigação e modelação nas Ciências Exactas e nas Ciências Sociais. Apesar de, com frequência, quer a variável dependente quer as variáveis independentes serem medidas com erros e apresentarem variabilidade natural, os modelos de tipo II, que considerem este tipo de erros, são raramente utilizados. Neste estudo avaliou-se, via simulação de Monte-Carlo, a consistência e eficiência de três modelos de tipo II (Eixo Principal reduzido, Método robusto de Kendall, Método dos 3 grupos de Bartlett) e o Método da Máxima Verosimilhança com rácio de variâncias-dos-erros conhecido, comparativamente ao método de tipo I dos mínimos quadrados comuns. Quando a variável independente é medida com erro, as estimativas do declive obtidas pelo método dos mínimos quadrados são enviesadas para 0, e o mesmo enviesamento foi observado com os métodos de Kendall e de Bartlett. O método do eixo principal reduzido produziu estimativas centradas no parâmetro a estimar quando a variabilidade da variável independente é da mesma magnitude da variabilidade da variável dependente. Contudo, este método produz sub- ou sobre estimativas do declive quando a variabilidade da variável independente é, respectivamente, superior ou inferior à variabilidade da variável dependente. Finalmente, o método da máxima verosimilhança com rácio de variâncias conhecido produz estimativas do declive erróneas para amostras de pequena dimensão, convergindo para os valores esperados de forma eficiente e consistente, quando as dimensões das amostras são superiores a 30.

Palavras–chave: Regressão linear com erros-nas-variáveis; modelos lineares de tipo II; Mínimos quadrados, Máxima verosimilhança; Eixo Principal Reduzido; Método de Bartlett; Método Robusto de Kendall.

Abstract: Type I regression models that allow for measurement error in the criterion only are used ubiquitously in the exact and social sciences. Although, more often than not, the predictor variable is also affected by natural variability and measurement errors, type II or error-in-variables regression models that account for these errors are seldom used. This is due in part to a lack of knowledge on the consistency and efficiency of type II estimators. In this paper, I present Monte-Carlo simulations, to study the consistency and efficiency of 3 type II linear regression models (Bartlett's 3 group method; Reduced Principal Axis and Kendall Robust method) as compared with maximum likelihood with errors-variance ratio known, and ordinary least squares. When

only the criterion variable is affected by measurement errors, the methods evaluated, with the exception of the Reduced Principal Axis, gave consistent estimates with OLS showing the highest efficiency. When errors in the predictor are as large as the errors in the criterion, only the Reduced Principal Axis and the Maximum Likelihood methods produced consistent and efficient estimates for the slope and the intercept. The other methods produced estimates for the slope biased towards 0. Finally, when the errors in the predictor were larger than the errors in the criterion (in this study, twice as large), only Maximum likelihood produced consistent estimates for large sample sizes (30 or more). The Reduced principal axis produced under- or over-estimates of the true positive slope when the errors in the predictors were larger or smaller, respectively, than the errors in the criterion. If the slope being estimated was negative, than one obtains over- and underestimates whenever the errors in X were larger or smaller than errors in Y respectively. For smaller sample sizes, Maximum Likelihood behaved erroneously, with large variance.

Keywords: Errors-in-variables; Type II Linear Regression; Reduced Principal Axis, Bartlett 3 groups method; Kendall Robust method; Maximum Likelihood with known errors-variance ratio; Ordinary Least Squares.

1 Introdução

O modelo de regressão linear simples é uma das mais interessantes ferramentas de análise estatística ao dispor do investigador das Ciências Exactas, Sociais e Biológicas. Este tipo de modelos, permite estabelecer relações funcionais entre uma variável dita dependente $(Y_i)(i=1,\ldots,n)$ e uma variável independente ou predictora (X_i) do tipo:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

Neste modelo, a variação em Y é explicada por um valor constante (β_0 - ordenada na origem), por uma influência linear e aditiva de X traduzida pelo coeficiente de regressão (β_1 - declive da recta) e por uma componente de erro aleatório (ε_i) de medida e/ou variação natural em Y. Inferência sobre β_1 no modelo (1) permite concluir sobre a significância da relação funcional de Y em X, e permite também usar o modelo com fins preditivos, *i.e.* estimar o valor médio ou valor esperado de Y_i em função dos valores que X_i toma:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{2}$$

Para que a inferência sobre o modelo (1) e a sua utilização com fins preditivos seja válida,é necessário validar um conjunto de pressupostos. É usual considerar que i) apenas a variável dependente é afectada de erros de medição, registo ou variação natural e que ii) os erros são independentes e com distribuição normal de média zero e desvio-padrão constante [*i.e.* $\varepsilon \sim IIN(0,\sigma)$]. Quando estas condições de aplicação são válidas o modelo de regressão linear simples diz-se

de Tipo I e os coeficientes do modelo são estimáveis de forma consistente e eficiente pelo método dos Mínimos Quadrados Comuns. Este tipo de modelo é usado regularmente em contextos de investigação aplicada à Psicologia, Ciências Sociais, Ciências Exactas e Biológicas. Porém, na maior partes dos cenários de investigação, nomeadamente nos estudos correlacionais, a variável independente é medida e não manipulada estando sujeita a erros de medição, para além de apresentar variação natural não controlada pelo investigador. Enquanto que nas ciências exactas a calibração frequente dos instrumentos de medida assegura que as variáveis em estudo são medidas com erro negligível, nas ciências sociais e humanas muitas das variáveis em estudo só podem ser medidas de forma indirecta e com erro. Assim, o pressuposto de que X é livre de erro raramente é valido (ver *e.g.* Sokal e Rolf, 1995) o que faz com que o modelo de tipo I não possa ser utilizado com fins inferenciais. Neste cenário, as estimativas de β_1 obtidas pelo método dos Minímos Quadrados Comuns são enviesadas para 0 e o teste à significância estatística de β_1 apresenta uma probabilidade de erro de tipo II acrescida (Riggs et al, 1978; Snedecor e Cochran, 1989, Fuller, 1987; Cheng e van Ness, 1999). Por outro lado, se os erros cometidos na medição de Y não forem $IIN(0,\sigma)$, as propriedades distribucionais do modelo não são válidas, o mesmo acontecendo à inferência realizada em torno da significância da relação funcional.

Os modelos de regressão linear em que quer a variável dependente quer a variável independente são aleatórias e com distribuição normal bivariada, são designados por 'modelos de tipo II' (Sokal e Rolf, 1995) ou por 'modelos de erros-nas-variáveis' (Fuller, 1987; Chang e van Ness, 1999). Este modelo assume que as verdadeiras variáveis latentes η_i e ξ_i (i=1,...,n) estão relacionadas por:

$$\eta_i = \beta_0 + \beta_1 \xi_i \tag{3}$$

Porém, (η_i, ξ_i) não são directamente observáveis sendo apenas possivel observar (Y_i, X_i), que são a operacionalização das variáveis latentes afectadas de erros de mensuração $(\varepsilon_i, \delta_i)$ aditivos:

$$Y_i = \eta_i + \varepsilon_i \tag{4}$$

$$X_i = \xi_i + \delta_i \tag{5}$$

onde ε_i e δ_i são variáveis aleatórias independentes com distribuição normal de média 0 e variâncias constantes *i.e.* $\varepsilon \sim N(0, \sigma_\varepsilon)$ e $\delta \sim N(0, \sigma_\delta)$ sendo $Cov(\delta, \varepsilon) = 0$.

Se bem que os modelos de regressão linear de tipo II não sejam novos em diversas áreas de estudo (nomeadamente na Biologia, ver *e.g.* Riggs *et al*, 1978), noutras, como por exemplo na Psicologia e Ciências Sociais, e de um modo geral, não se encontram menções à utilização destes modelos (uma pesquisa na base de dados PsycARTICLES em 01/05/2006 usando como palavras-chave 'model II regression' ou 'errors-in-variables' retornou apenas dois 'matches' para os últimos 10 anos: Friedman et al. 2004; Schuster, 2004). Mesmo nas Ciências

Biológicas e Econométricas, a utilização destes modelos, ou o reconhecimento de que os predictores são medidos com erro é, surpreendentemente, pouco referenciada (Quinn e Keough, 2002). A utilização dos modelos de tipo II tem sido alvo de alguma investigação e controvérsia, nomeadamente na sua utilização em Ciências Biológicas, sendo difícil apresentar recomendações simples sobre qual dos diferentes modelos de tipo II se deve utilizar. Esta ausência de consenso resulta em parte, da falta de informação quanto aos melhores estimadores para os coeficientes de regressão dos modelos de tipo II. Após uma revisão limitada da bibliografia, Maroco (2003) sugere que 3 métodos de estimação de modelos de tipo II (nomeadamente o Método do Eixo Principal Reduzido, Método Robusto de Kendall e o Método dos 3 grupos de Bartlett) podem ser utilizadas com alguma simplicidade e em alternativa ao método dos mínimos quadrados (utilizado nos modelos de tipo I) para estimar os coeficientes de regressão. Nesta comunicação, avaliou-se por intermédio de simulações de Monte-Carlo, a consistência e eficiência destes estimadores e dos estimadores de máxima verosimilhança com rácio de variâncias-dos-erros conhecido.

2 Simulação e modelos de regressão

2.1 Simulação de Monte-Carlo

Valores de ξ (valores não-observados da variável latente predictora) foram gerados com a função *RndNormal*(1) implementada no software STATISTICA 7 (StatSoft, Tulsa, OK) com amplitude de -6 a 6 para 95% dos valores. Os valores da variável latente critério η (valores reais, não-observados) foram obtidos por $\eta = 1+1\xi$. Erros aleatórios com distribuição normal de média 0 e variância constante foram posteriormente adicionados às variáveis latentes, para reproduzir erros de medição e variabilidade natural não-controlada nas variávies (Y, X) observadas, como descrito pelas equações (3) e (4). Recorrendo a simulação de Monte-Carlo, foram geradas 10000 amostras com dimensões (n) entre 4 e 400 e com diferentes valores de σ_δ e σ_ε. Os pares (Y_i, X_i) $(i=1,\ldots,n)$ foram depois usados para estimar β_1 e β_0 usando os estimadores descritos a seguir.

2.2 Estimadores do declive e ordenada na origem para variáveis medidas com erro

No método dos **Mínimos Quadrados Comuns** (MQ), os estimadores do declive e ordenada na origem são, respectivamente:

$$\beta_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} \tag{6}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \tag{7}$$

No **Método do Eixo Principal Reduzido** (EPR), ou da média geométrica, os estimadores usados são:

$$\beta_{1EPR} = Sign[Cov(X,Y)]\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = Sign[Cov(X,Y)]\sqrt{\frac{S_{YY}}{S_{XX}}} \quad (8)$$

$$\beta_{0EPR} = \bar{Y} - \beta_{1EPR}\bar{X} \quad (9)$$

onde a função $Sign(.)$ retorna o sinal (+ ou −) do seu argumento.

No **Método Robusto de Kendall** (Ken) os valores (Y_i, X_i) (i=1,...,n) são ordenados por ordem crescente de X_i e calculam-se $n(n-1)/2$ declives para os valores adjacentes:

$$S_{j+1,j} = \frac{Y_{j+1} - Y_j}{X_{j+1} - X_j} \quad (10)$$

O declive é estimado pela mediana de $S_{j+1,j}(j = 1, \ldots, n-1)$ e a ordenada na origem é a mediana de $a_i = Y_i - \tilde{b}X_i$ $(i = 1, \ldots, n)$.

No **Método dos 3 grupos de Bartlett** (Bart) as observações (Y_i, X_i) (i=1,...,n), depois de ordenadas por ordem crescente de X_i, são divididas em 3 grupos iguais, e o declive e a ordenada na origem são estimados por:

$$\beta_{1B} = \frac{\bar{Y_3} - \bar{Y_1}}{\bar{X_3} - \bar{X_1}} \quad (11)$$

$$\beta_{0B} = \bar{Y} - \beta_{1B}\bar{X} \quad (12)$$

No **Método da Máxima Verosimilhança** (MV) com rácio de variâncias-dos-erros $\lambda = \sigma_\varepsilon/\sigma_\delta$ conhecido, os estimadores são:

$$\beta_{1MV} = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{XX} - \lambda S_{YY})^2 + 4S_{XY}^2}}{2S_{XY}} \quad (13)$$

$$\beta_{0MV} = \bar{Y} - \beta_{1MV}\bar{X} \quad (14)$$

(Estes estimadores são também os estimadores de mínimos quadrados modificados de Cheng e Van Ness (1999)).

3 Resultados

Os valores médios das estimativas do declive e ordenada na origem, bem como os percentis 2.5 e 97.5 que definem um I.C. não paramétrico a 95% são reproduzidos nas figuras 1-3 em função da dimensão das amostras e das variâncias dos erros-de-medida usadas na simulação.

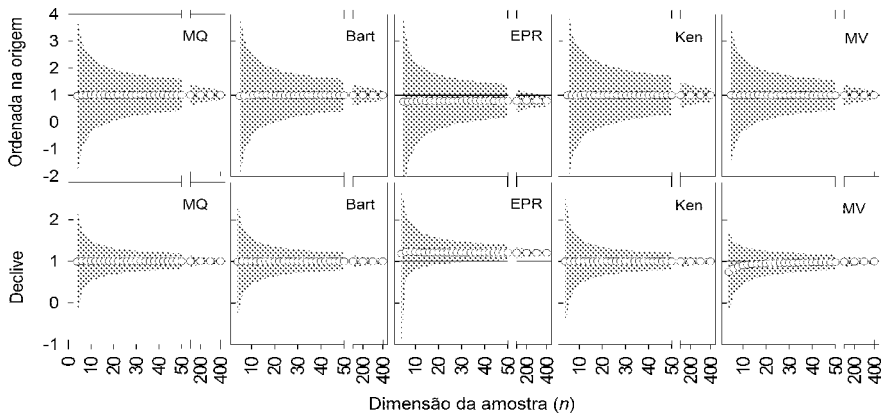

Figura 1: Estimativas da Ordenada na origem e do declive obtidas em 10000 amostras com dimensão (n) de 4 a 400 com os estimadores dos mínimos quadrados (MQ), 3 grupos de Bartlett (Bart), Eixo Principal reduzido (EPR), método robusto de Kendall (Ken) e máxima verosimilhança (MV) ajustados a dados (Y, X) gerados como $Y = \eta + \varepsilon$ com $\varepsilon \sim N(0, 2)$, $X = \xi + \delta$ com $\delta \sim N(0, 0)$ e $\eta = 1 + 1\xi$ como descrito na secção 'Simulação de Monte Carlo'. Os círculos representam a média das estimativas com intervalo percentilico a 95% (área a cinzento). No método MV usou-se $\delta \sim N(0, 0.001)$.

Quando a variável predictora é medida sem erro, todos os métodos testados, com a excepção do método do EPR, produzem estimativas centradas no verdadeiro valor populacional, sendo o método dos MQ o que apresenta estimativas com menor variância (Fig. 1).

Quando o erro-de-medida da variável predictora é da mesma ordem de magnitude do erro de medida na variável critério, apenas os métodos do EPR e de MV produzem estimativas centradas no parâmetro a estimar, enquanto que os restantes métodos testados tendem a sub-estimar o verdadeiro declive populacional e sobre-estimar a ordenada na origem (Fig. 2).

Quando o erro-de-medida da variável predictora é superior (neste estudo, 2x superior) ao erro de medida da variável critério apenas o método de MV produz estimativas consistentes ainda que para amostras de pequena dimensão as estimativas sejam erróneas e de variância elevada. Todos os outros métodos sub-estimam o verdadeiro declive e sobre-estimam a ordenada na origem (Fig. 3). Se os erros da variável predictora forem inferiores aos erros de medida da variável criterio, os modelos sobre-estimam o declive e sub-estimam a ordenada na origem (Resultados não apresentados).

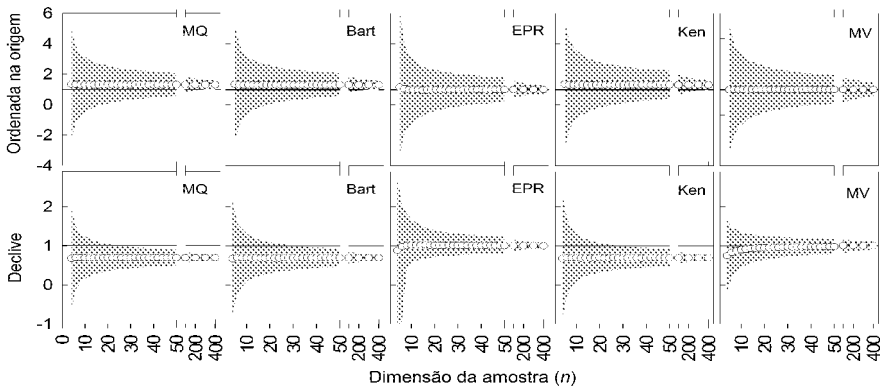

Figura 2: Estimativas da Ordenada na origem e do declive obtidas em 10000 amostras com dimensão (n) de 4 a 400 com os estimadores dos mínimos quadrados (MQ), 3 grupos de Bartlett (Bart), Eixo Principal Reduzido (EPR), método robusto de Kendall (Ken) e Máxima Verosimilhança (MV) ajustados a dados (Y, X) gerados como $Y = \eta + \varepsilon$ com $\varepsilon \sim N(0, 2)$, $X = \xi + \delta$ com $\delta \sim N(0, 2)$ e $\eta = 1 + 1\xi$ como descrito na secção 'Simulação de Monte Carlo'. Os círculos representam a média das estimativas com intervalo percentilico a 95% (área a cinzento).

4 Discussão

Se a variável predictora é medida sem erro, os estimadores dos mínimos quadrados comuns (MQ) são eficientes e consistentes, o mesmo acontecendo com o método de Bartlett (Bart), Kendall (Ken) e de Máxima verosimilhança (MV). Neste cenário, os estimadores do método de Kendall são os menos eficientes, enquanto que o Eixo Principal Reduzido (EPR) sobre-estima o verdadeiro declive. Quando o erro de medida da variável predictora é da mesma ordem de magnitude do erro de medida da variável critério, os estimadores do declive dos MQ, Bart e Ken produzem estimativas enviesadas para 0 (subestimando ou sobre-estimando o declive consoante o verdadeiro declive é positivo ou negativo, respectivamente), mas os estimadores do EPR e da MV são consistentes e eficientes. Finalmente, quando os erros de medida da variável predictora são superiores (inferiores) aos erros de medida da variável critério, nenhum dos estimadores dos modelos de tipo II é centrado e apenas a MV produz estimativas consistentes. Contudo, para amostras de pequena dimensão os estimadores de MV produzem estimativas erróneas e, apenas para amostra de dimensão considerável (30-40) ou mesmo muito grandes (200-400), os estimadores são eficientes.

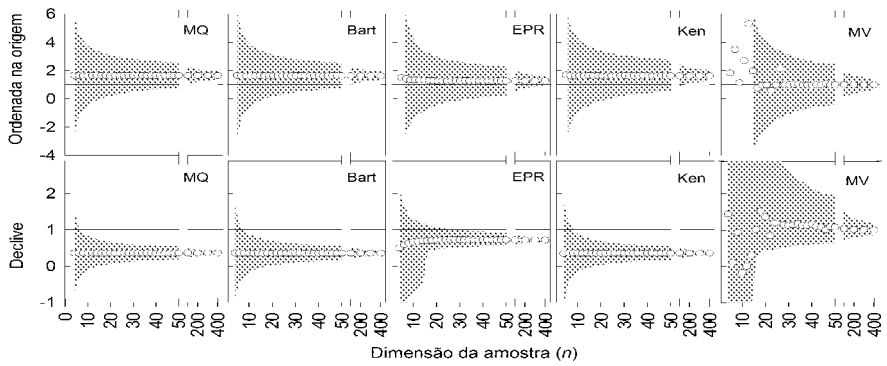

Figura 3: Estimativas da Ordenada na origem e do declive obtidas em 10000 amostras com dimensão (n) de 4 a 400 com os estimadores dos mínimos quadrados (MQ), 3 grupos de Bartlett (Bart), Eixo Principal Reduzido (EPR), método robusto de Kendall (Ken) e Máxima Verosimilhança (MV) ajustados a dados (Y, X) gerados como $Y = \eta + \varepsilon$ com $\varepsilon \sim N(0, 2)$, $X = \xi + \delta$ com $\delta \sim N(0, 4)$ e $\eta = 1 + 1\xi$ como descrito na secção 'Simulação de Monte Carlo'. Os círculos representam a média das estimativas com intervalo percentilico a 95% (área a cinzento).

Adicionalmente, o método da MV exige o conhecimento do rácio de variâncias dos erros de medida, exigindo medições replicadas, ou um conhecimento prévio do fenómeno sobre estudo. Em estudos exploratórios este conhecimento nem sempre está disponível, e este facto limita a aplicação dos estimadores de MV.

Referências

[1] Cheng, C-L. e Van Ness, J. W. (1999). *Statistical Regression with measurement error.* Arnold. London.

[2] Freedman, L. S. et al. (2004) A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* 60:172-181

[3] Fuller, WA (1987). *Measurement Error Models.* John Wiley. New York

[4] Maroco, J. (2003) *Análise Estatística com utilização do SPSS.* Editora Silabo, Lisboa

[5] Quinn, G. P., e Keough, M. J. (2002). *Experimental Data Analysis for Biologists.* Cambridge University Press. Cambridge.

[6] Riggs, D.S., et al (1978) Fitting straight lines when both variables are subject to error. *Life Sciences* 22:1305-1360.

[7] Schuster, C. (2004). How measurement error in dichotomous predictors affects the analysis of continuous criteria. *Psychology Science*, 46(1): 128-136.

[8] Sokal, R. R. e Rohlf, F. J. (1995) *Biometry. The principles and practice of Statistics in Biological Research.* 3rd ed. W. H. Freeman and Company. New York.