ESTIMAÇÃO E INFLUÊNCIA LOCAL NO MODELO DE REGRESSÃO SIMPLES COM ERRO NA VARIÁVEL USANDO A DISTRIBUIÇÃO NORMAL- CONTAMINADA

Lidiane da Silva Lima (lidinhalima@hotmail.com)

Prof. Dr. Victor H. Lachos (hlachos@ime.unicamp.br)

IMEEC – UNICAMP – CNPQ/PIBIC

Palavras-chave: EM-algorithm, local influence, contaminated normal.

Introdução

Assumir que as observações seguem uma distribuição Normal é uma suposição rotineira em modelos lineares. No entanto, esta suposição pode não ser realista ocultando importantes características da variação que está presente nos dados. A distribuição Normal Contaminada (Little, 1989) é atrativa porque permite modelar dados com certa proporção de outliers.

Metodologia

O modelo com erro de medida é dado da seguinte forma:

$$X_i = x_i + u_i$$

$$Y_i = \alpha + \beta x_i + e_i$$

onde tanto os erros aleatórios (u_i e e_i) como a variável latente x (não observável) seguem conjuntamente uma distribuição Normal Contaminada.

Uma v.a. C com distribuição Normal Contaminada pode ser construída a partir de uma v.a. Normal Z, com vetor de médias **0** e matriz de variância-covariância $\boldsymbol{\Sigma}$, e uma v.a. de mistura discreta U dada pela seguinte expressão

$$\mathbf{C} = \boldsymbol{\mu} + K^{1/2}(U)\,\mathbf{Z},$$

onde $\boldsymbol{\mu}$ é um vetor de locação e $K^{1/2}(U)$ é uma função da v.a. U. Neste trabalho considera-se $K^{1/2}(U)=1/u$. A função de probabilidade de U é denotada por

$$h(u) = \nu I_{(u=\gamma)} + (1-\nu) I_{(u=1)}$$

onde ν é a proporção de outliers e γ é um fator de escala.

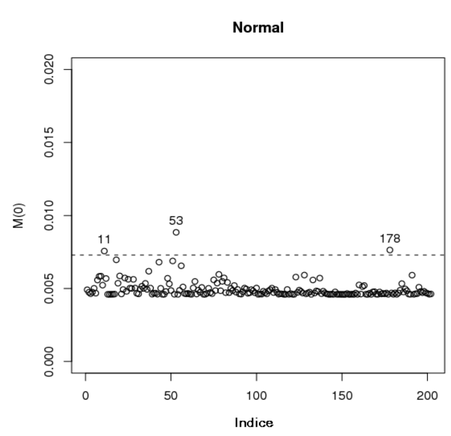
Para a estimação dos parâmetros usa-se o algoritmo EM, que divide-se em duas partes: cálculo da Q-function (passo E) e sua maximização com respeito aos parâmetros da regressão (passo M).

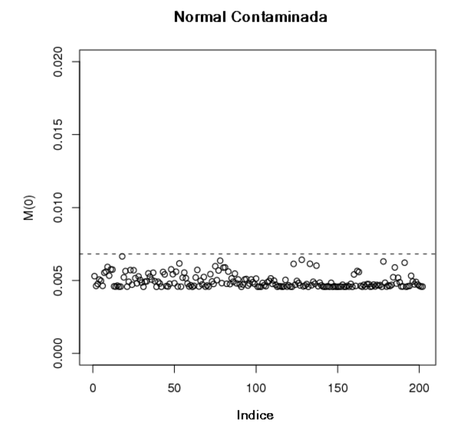
A fim de detectar observações que, sob menores perturbações do modelo, exercem grande influência sobre as estimativas dos parâmetros, utilizou-se a abordagem de influência local de Zhu e Lee (2001) e alguns tipos de perturbação.

Resultados

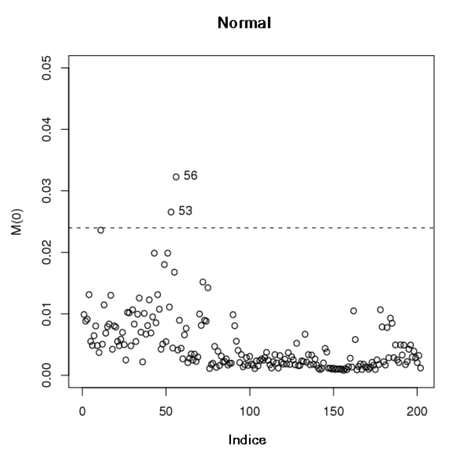
Como aplicação da teoria apresentada, usou-se o conjunto de dados AIS (disponível no software R, pacote SN). Foi possível verificar a redução de pontos influentes quando a distribuição Normal foi substituída pela distribuição Normal Contaminada.

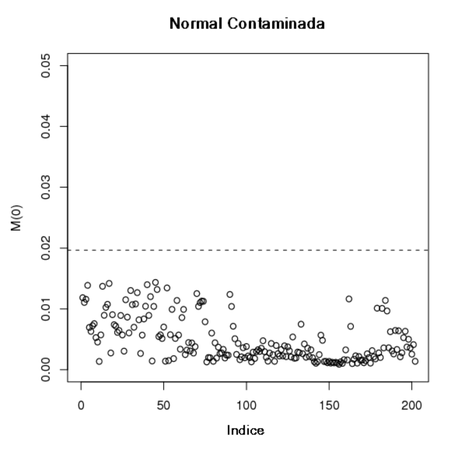
Perturbação da razão de variância: detectar observações que exercem influência na suposição de homocedasticidade (variância constante).





Perturbação da variável resposta: detectar observações que exercem influência nos valores preditos.





Conclusões

Os dados pertencem a uma distribuição mais robusta que a Normal, pois a influência de pontos discrepantes é reduzida quando utilizou-se a distribuição Normal Contaminada.

Bibliografia

[1] Lange, K. L., Little, J. A. and Taylor, M. G. J. (1989). Robust statistical modeling using the t distribution. Journal of the American Statistical Association, 84, 881-896.

[2] Zhu, H. and Lee, S. (2001). Local influence for incomplete-data models. Journal of the Royal Statistical Society, Series B, 63, 111-126.