

RT-MAE 2007-13

**ALGUMAS MEDIDAS DE DIAGNÓSTICO
NO MODELO DE REGRESSÃO
LOGÍSTICA DE BOX-COX**

**Noemi Ichihara Ishikawa
and
Sílvia Nagib Elian**

Palavras-Chave: Modelo de Regressão Logística de Box-Cox, Diagnóstico, Diferença de Verossimilhanças, Análise de Influência Parcial.

Classificação 62J20

- Julho de 2007 -

Algumas Medidas de Diagnóstico no Modelo de Regressão Logística de Box-Cox

Noemi Ichihara Ishikawa¹

Silvia Nagib Elian²

¹ E-mail: noemi.ichihara@gmail.com

² Departamento de Estatística
Instituto de Matemática e Estatística,
Universidade de São Paulo
C. P. 66281 – São Paulo, SP Brasil
E-mail: selian@ime.usp.br

Resumo

Transformações em covariáveis são aplicadas em análise de regressão por vários motivos e famílias de transformações paramétricas, como a de Box-Cox, são muito comuns. Quando uma transformação paramétrica é usada, o estimador de máxima verossimilhança do parâmetro da transformação é frequentemente sensível a pequenas perturbações dos dados. A maioria dos métodos de diagnóstico nessa linha trata do uso de transformações na variável resposta ou em ambas, variável resposta e covariável. Além disso, o estudo restringe-se a modelos lineares.

Neste trabalho, descreveremos algumas técnicas de diagnóstico quando são utilizadas transformações na covariável em modelos lineares generalizados, com aplicação a um conjunto de dados com resposta binária, ao qual foi ajustado o modelo de regressão logística de Box-Cox.

Palavras-Chave: Modelo de Regressão Logística de Box-Cox, Diagnóstico, Diferença de Verossimilhanças, Análise de Influência Parcial.

1. Introdução

O modelo de regressão logística de Box-Cox foi proposto por Guerrero e Johnson (1982), que aplicaram a transformação de Box-Cox

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \log x & , \lambda = 0. \end{cases}$$

à variável explicativa do modelo de regressão logística.

Nessas condições, se Y é uma variável resposta binária, com valores 0 ou 1 e admite-se que $p = P(Y=1)$ é função de uma variável explicativa, também denominada covariável, X , o modelo de regressão logística de Box-Cox é dado por

$$\log \left[\frac{p}{1-p} \right] = \beta_0 + \beta_1 x^{(\lambda)}$$

que equivale à expressão

$$p = \frac{\exp(\beta_0 + \beta_1 x^{(\lambda)})}{1 + \exp(\beta_0 + \beta_1 x^{(\lambda)})},$$

com $x > 0$.

De acordo com essa construção, o modelo de regressão logística seria apropriado após uma transformação da covariável, mas não sabemos antecipadamente qual seria ela. O parâmetro de transformação λ deve ser estimado, e uma possibilidade é a utilização do método de máxima verossimilhança, que estimaria conjuntamente β_0 , β_1 e λ .

Com esse objetivo, Siqueira(1992) sugere o seguinte procedimento de estimação em dois estágios:

Inicialmente, fixado um particular valor de λ , obtém-se as estimativas de máxima verossimilhança de β_0 e β_1 através do procedimento de Newton-Raphson. Calcula-se então o logaritmo da função de verossimilhança com base no valor de λ fixado e $\hat{\beta}_0$, $\hat{\beta}_1$, estimadores de máxima verossimilhança obtidos. Repete-se o procedimento para cada valor de λ pertencente a um particular conjunto de possíveis valores. Sugere-se que λ seja escolhido no intervalo $[-2, 2]$, que pode ser ampliado se necessário. A estimativa de máxima verossimilhança de λ é o valor desse parâmetro que maximiza o logaritmo da função de verossimilhança.

Verifica-se que esse método de estimação, embora bastante simples, geralmente é executado com sucesso.

Siqueira e Taylor (1999) analisam o custo de estimação de λ no modelo de regressão logística de Box-Cox, avaliando a inflação da variância dos estimadores de alguns parâmetros de interesse devido à adição de λ .

Adicionalmente, Lee e Yick (1999) destacam a alta sensibilidade do estimador de tal parâmetro a pequenas perturbações nos dados. A maioria dos métodos de diagnóstico neste caso trata do uso de transformações na variável resposta ou em ambas, variável resposta e explicativa e restringem-se a modelos lineares. Os autores desenvolvem técnicas de diagnóstico quando são utilizadas transformações na variável explicativa em modelos lineares generalizados.

O objetivo do presente trabalho é apresentar uma breve análise de diagnóstico no modelo de regressão logística de Box-Cox.

Serão descritos inicialmente dois métodos de diagnóstico propostos por Lee e Yick (1999) para o modelo linear generalizado com transformação na variável explicativa.

Posteriormente, será feita uma adaptação desses métodos ao modelo de regressão logística de Box-Cox através de uma aplicação a um conjunto de dados com resposta binária.

2. Métodos de Diagnóstico no Modelo Linear Generalizado com transformação na variável explicativa

Serão apresentados nessa seção os métodos de diagnóstico: diferença de verossimilhanças e influência parcial para o modelo linear generalizado.

Diferença de verossimilhanças

Seja $y = (y_1, \dots, y_n)^T$ o vetor de respostas com densidade:

$$f_y(y_i; \theta) = \exp \{ [y_i \theta_i - b(\theta_i)] / a(\phi) + c(y_i, \phi) \}$$

com $\theta_i = k(\eta_i)$, em que η_i é o preditor linear e $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ são funções conhecidas. Sem perda de generalidade, o parâmetro de dispersão ϕ é supostamente conhecido ou é substituído por um estimador $\hat{\phi}$ e escrito como $\hat{a} = a(\hat{\phi})$, que fornece uma densidade da família exponencial com parâmetro natural θ . O logaritmo da função de verossimilhança é então dada por:

$$\hat{a}^{-1} \sum_{i=1}^n [y_i k(\eta_i) - b\{k(\eta_i)\}].$$

Consideremos X uma matriz de covariáveis do modelo, $n \times (p + q)$, particionada na forma:

$$X = [x, z] = [x_1 \dots x_p, z_1 \dots z_q],$$

com $x_1 \dots x_p, z_1 \dots z_q$ correspondentes às colunas da matriz X .

Um melhor ajuste para o modelo linear generalizado proposto pode ser frequentemente obtido transformando uma ou mais das covariáveis z presentes em $X = [x, z]$.

Seja o preditor linear do modelo transformado:

$$\eta = x\delta + G(z, \lambda)\xi,$$

em que $G(z, \lambda) = [g_1(z_{(1)}, \lambda_1), \dots, g_q(z_{(q)}, \lambda_q)]$ é uma matriz $n \times q$ e

$$g_j(z_{(j)}, \lambda_j) = \langle g_j(z_{(1j)}, \lambda_j), \dots, g_j(z_{(nj)}, \lambda_j) \rangle^T$$

representa uma família de transformação conhecida, duplamente diferenciável, indexada por λ_j ($j = 1, \dots, q$). e δ e ξ são vetores de parâmetros, p e q dimensionais. Temos aqui um interesse especial no vetor de parâmetros λ .

Sejam $L(\lambda, \delta, \xi)$ o logaritmo da função de verossimilhança, $\bar{\delta}(\lambda)$ e $\bar{\xi}(\lambda)$ funções que maximizam $L(\lambda, \delta, \xi)$, para λ fixo e $L(\lambda; \bar{\delta}(\lambda), \bar{\xi}(\lambda))$ o correspondente máximo obtido. Para avaliar a influência de casos individuais na estimativa de máxima verossimilhança $\hat{\lambda}$ de λ , a diferença entre $\hat{\lambda}$ e $\hat{\lambda}_{(i)}$, a estimativa de máxima verossimilhança sem o caso i , pode ser mensurada através da medida diferença de verossimilhanças dada por:

$$LD_i = 2[L(\hat{\lambda}) - L(\hat{\lambda}_{(i)})],$$

em que $L(\lambda) = L(\lambda; \bar{\delta}(\lambda), \bar{\xi}(\lambda))$.

Para o caso especial de regressão linear, Wei e Hickernell (1996) (citado por Lee e Yick (1999)) obtiveram aproximações para simplificar os cálculos dos valores de LD_i . Um grande valor LD_i indica que provavelmente a estimativa $\hat{\lambda}$ é fortemente dependente do caso i .

Análise de influência parcial

Com a notação já introduzida, consideremos novamente o preditor linear antes da transformação da covariável z , dado por

$$\eta_0 = x\beta + z\gamma. \quad (2.1)$$

O modelo após a transformação é:

$$\eta = x\beta + G(z, \lambda)\xi \quad (2.2)$$

de modo que $z = G(z, \lambda_0)$ representaria a inexistência de transformação.

Um teste de hipótese $H_0: \lambda = \lambda_0$ pode ser baseado em $D_0 - D$, a redução da função desvio do modelo (2.1) para o modelo (2.2). É importante verificar se qualquer observação tem um impacto indevido neste teste. De modo geral, a função desvio de um particular modelo é definida como $D = -2 \ln \left[\frac{L}{L_0} \right]$, em que L é a função de verossimilhança desse modelo calculada nos estimadores de máxima verossimilhança e L_0 é a correspondente função para o modelo saturado. Denotando-se $D_{0[i]}$ e $D_{[i]}$ as funções desvio do modelo (2.1) e (2.2) respectivamente após excluir o caso i , Lee (1988) (citado por Lee e Yick (1999)) sugeriu uma medida de influência parcial para o impacto do caso i na transformação como:

$$d_i = (D_0 - D) - (D_{0[i]} - D_{[i]}),$$

que compara reduções no desvio devido à transformação com e sem a i -ésima observação. Novamente aqui, grandes valores de d_i implicam em grande alteração da redução no desvio com a retirada da i -ésima observação.

3. Adaptação ao Modelo de Regressão Logística de Box-Cox

Neter, Kutner, Nachtsheim e Wasserman (1996) apresentam um conjunto de dados relativo à variável resposta compra de carro (sim ou não) em função de algumas características da família e que será analisado a seguir.

Uma montadora de veículos contrata uma empresa de pesquisa de mercado na condução de um estudo piloto para analisar a probabilidade de uma família comprar um carro novo durante o ano seguinte. Uma amostra casual de 33 famílias foi selecionada e foram coletados dados da renda familiar anual (X , em milhares de dólares). Doze meses depois, foi conduzida uma entrevista para determinar se, no decorrer do ano, houve compra de um veículo novo ($Y = 1$) ou não ($Y = 0$). Estes dados são apresentados na Tabela 3.1.

Tabela 3.1 – Dados relativos às 33 famílias.

Família (i)	Y_i	x_i
1	0	32
2	0	45
3	1	60
4	0	53
5	0	25
6	1	68
7	1	82
8	1	38
9	0	67
10	1	92
11	1	72
12	0	21
13	0	26
14	1	40
15	0	33
16	0	45
17	1	61
18	0	16
19	1	18
20	0	22
21	0	27
22	1	35
23	1	40
24	0	10
25	0	24
26	1	15
27	0	23
28	0	19
29	1	22
30	0	61
31	0	21
32	1	32
33	0	17

Com base nos dados apresentados na Tabela 3.1, serão feitas as análises de diagnóstico apresentadas na seção anterior. Será ajustado o modelo $\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x^{(\lambda)}$. Para estimar seus parâmetros, conforme descrito na seção anterior, serão fixados valores de λ no intervalo $[-2, 2]$: -2; -1,5; -1; -0,5; 0,5; 1; 1,5; 2 e para cada um deles, obteremos as estimativas de máxima verossimilhança de β_0 e β_1 . O logaritmo da função de verossimilhança desse modelo, indexado pelos parâmetros $\theta = (\beta_0, \beta_1, \lambda)$ é dado por

$$\begin{aligned} L &= \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \\ &= \sum_{i=1}^n y_i (\beta_0 + \beta_1 x_i^{(\lambda)}) - \sum_{i=1}^n \log(1 + \exp(\beta_0 + \beta_1 x_i^{(\lambda)})) \end{aligned}$$

com $p_i = p(x_i)$.

Calcula-se então o logaritmo da função de verossimilhança com base no valor de λ fixado e $\hat{\beta}_0, \hat{\beta}_1$, estimativas de máxima verossimilhança obtidas para cada um dos valores de λ . A estimativa de máxima verossimilhança de λ é o valor desse parâmetro que maximiza o logaritmo da função de verossimilhança.

Foi utilizado o programa estatístico R, para o ajuste do modelo com base nos dados.

O modelo ajustado é $\log\left[\frac{\hat{p}(x)}{1-\hat{p}(x)}\right] = -1,209212 + 0,000974x^{(2)}$.

Após esse procedimento, para efeito de diagnóstico, retira-se cada uma das observações e repete-se o cálculo feito anteriormente. Com isso, é possível calcular cada uma das estimativas de máxima verossimilhança de λ sem o caso i , o que permitirá o cálculo da diferença de verossimilhanças,

$$DV_i = 2[L(\hat{\lambda}) - L(\hat{\lambda}_{(i)})].$$

Adicionalmente, obtivemos a medida $DV_{(i)} = 2[L(\hat{\lambda}) - L_{(i)}(\hat{\lambda}_{(i)})]$, onde $L_{(i)}(\hat{\lambda}_{(i)})$ é a função de verossimilhança com base nas $n - 1$ observações restantes, calculada no ponto $\hat{\lambda}_{(i)}$.

A Tabela 3.2 apresenta as estimativas de máxima verossimilhança de λ quando retiradas cada uma das observações e os valores de DV_i e $DV_{(i)}$.

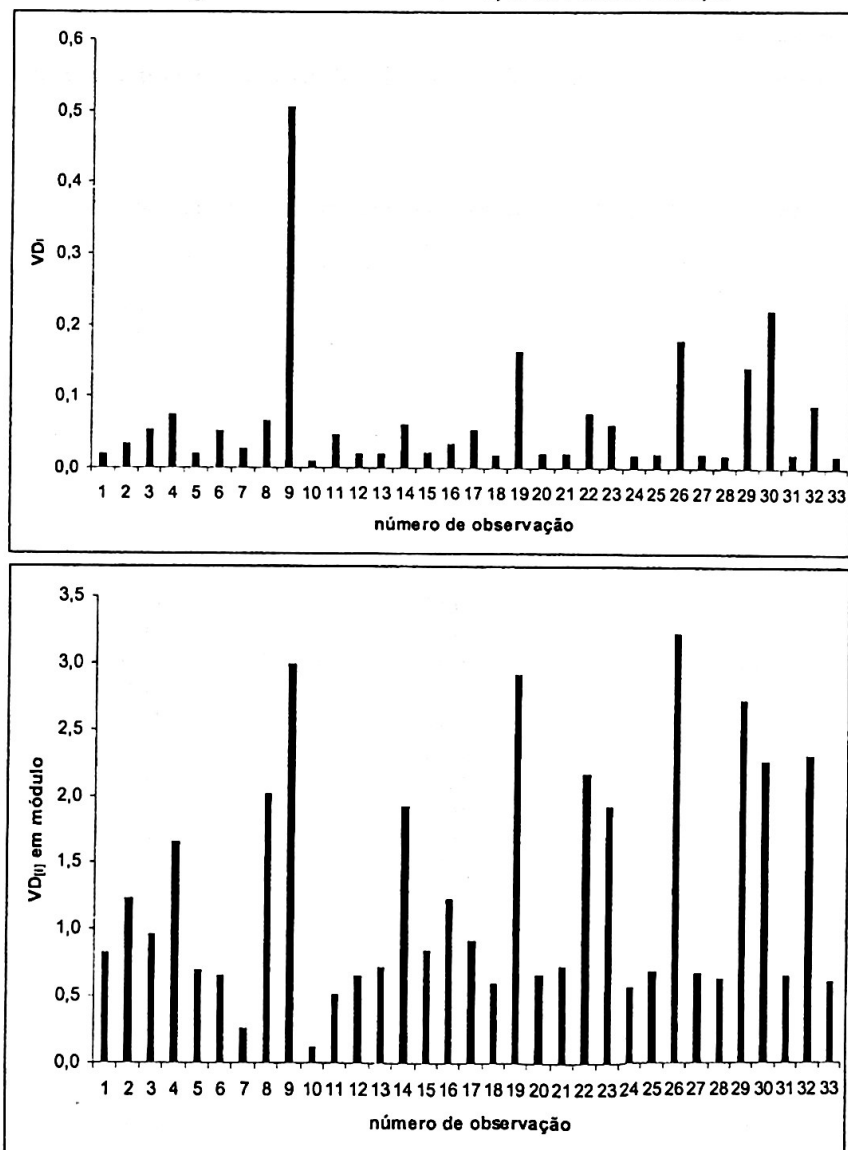
Como pode ser observado na Tabela 3.2, os valores mais influentes são relativos às famílias 9, 19, 26 e 30, quando o cálculo é feito por $2[L(\hat{\lambda}) - L(\hat{\lambda}_{(i)})]$ e para o cálculo de $2[L(\hat{\lambda}) - L_{(i)}(\hat{\lambda}_{(i)})]$, os dados influentes são das famílias 9, 19, 26 e 29. Observa-se que são famílias que compraram automóvel e têm renda familiar baixa (caso das famílias 19, 26 e 29) ou famílias que não compraram um automóvel e têm renda familiar alta (famílias 9 e 30).

Tabela 3.2 – Diferenças de verossimilhanças para os dados da Tabela 2.1.

Observação retirada	$\hat{\lambda}_{(i)}$	$\hat{\lambda} - \hat{\lambda}_{(i)}$	$2[L(\hat{\lambda}) - L(\hat{\lambda}_{(i)})]$	$2[L(\hat{\lambda}) - L_{(i)}(\hat{\lambda}_{(i)})]$
nenhuma	2			
1	2	0	0,020093	-0,81857
2	1,5	0,5	0,033290	-1,22788
3	2	0	0,052070	-0,96537
4	1,5	0,5	0,074718	-1,65244
5	2	0	0,018934	-0,69749
6	2	0	0,050348	-0,65042
7	2	0	0,026162	-0,26141
8	2	0	0,065512	-2,01869
9	2	0	0,503970	-2,98286
10	2	0	0,009216	-0,11395
11	2	0	0,045782	-0,51673
12	2	0	0,018598	-0,64713
13	2	0	0,019043	-0,71204
14	2	0	0,060222	-1,92038
15	2	0	0,020397	-0,83994
16	1,5	0,5	0,033290	-1,22788
17	2	0	0,052290	-0,92292
18	2	0	0,018283	-0,59972
19	1	1	0,162789	-2,91422
20	2	0	0,018663	-0,65861
21	2	0	0,019166	-0,72744
22	2	0	0,075315	-2,16363
23	2	0	0,060222	-1,92038
24	2	0	0,018018	-0,56212
25	2	0	0,018838	-0,68376
26	0,5	1,5	0,178257	-3,22293
27	2	0	0,018737	-0,67081
28	2	0	0,018456	-0,62624
29	2	0	0,140307	-2,71912
30	2	0	0,220395	-2,26286
31	2	0	0,018598	-0,64713
32	2	0	0,087764	-2,30409
33	2	0	0,018339	-0,60795

A Figura 3.1 apresenta os gráficos de DV_i e o módulo de $DV_{[i]}$ para as 33 observações.

Figura 3.1 - Gráficos das diferenças de verossimilhança.



Para o cálculo das medidas de influência parcial foram usadas as funções desvio das saídas do programa estatístico R, quando ajustados os modelos (2.1) e (2.2) com base nos dados da Tabela 3.1.

Os valores das diferenças nas funções desvio e da medida d_i estão apresentados na Tabela 3.3.

Tabela 3.3 – Medida de influência parcial com base nos dados da Tabela 3.1

Observação retirada	$D_{0[i]} - D_{[i]}$	$(D_0 - D) - (D_{0[i]} - D_{[i]})$
nenhuma		
1	0,13	0,08
2	0,04	1,17
3	0,27	-0,06
4	0,04	0,17
5	0,21	0,00
6	0,18	0,03
7	0,07	0,14
8	0,44	-0,23
9	0,37	-0,16
10	0,05	0,16
11	0,14	0,07
12	0,26	-0,05
13	0,19	0,02
14	0,44	-0,23
15	0,12	0,09
16	0,04	0,17
17	0,26	-0,05
18	0,31	-0,10
19	0,00	0,21
20	0,24	-0,03
21	0,19	0,02
22	0,41	-0,20
23	0,44	-0,23
24	0,37	-0,16
25	0,22	-0,01
26	0,05	0,16
27	0,23	-0,02
28	0,28	-0,07
29	0,12	0,09
30	0,13	0,08
31	0,26	-0,05
32	0,38	-0,17
33	0,30	-0,09

$D_0 - D = 0,21$

Observamos valores de d_i relativamente baixos em módulo, indicando inexistência de pontos influentes no teste que avalia a necessidade de transformação.

4. Considerações finais

Lee e Yick (1999), ainda no contexto de transformações no modelo linear generalizado, apresentam algumas medidas de influência local, que não foram aqui calculadas e seriam sugestões para um trabalho futuro.

Devido a facilidades computacionais e à ampla variedade de modelos existentes na literatura, parece ser uma tendência atual preferir utilizar um novo modelo, ao invés de se transformar os dados.

Acreditamos no entanto que o uso de transformações no modelo de regressão logística, apesar da riqueza de classes de modelos alternativos, ainda é uma ótima opção. Destacamos a facilidade de interpretação dos parâmetros desse modelo, aliados a sua simplicidade e popularidade e ainda, a disponibilidade na maioria dos pacotes estatísticos

Referências Bibliográficas

- Guerrero, V. M. and Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models, *Biometrika*, 69, 2, 309-314.
- Lee, A. H. (1988). Assessing partial influence in generalized linear models, *Biometrics*, 44, 71-77.
- Lee, A. H. and Yick, J. S. (1999). Covariate Transformation Diagnostics for Generalized Linear Models, *Ann. Inst. Statist. Math.*, 51, No. 2, 383-398.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*. Fourth edition. Chicago: Irwin. 1408p.
- Siqueira, A. L. (1992). *Box-Cox Transformation of the Covariate in Logistic Models*. Tese de doutorado, UCLA, 130p.
- Siqueira, A. L. and Taylor, J. M. G., (1999). Treatment Effects in a Logistic Model Involving the Box-Cox Transformation, *Journal of the American Statistical Association*, 94, No. 445, 240-246.

Wey, B. C. and Hickernell, F. J. (1996). Regression transformation diagnostics for explanatory variables, *Statist. Sinica*, 6, 433-454.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2007-01 - POLPO, A., PEREIRA, C.A.B. Reliability nonparametric bayesian estimation in parallel systems 2007. 12p. (RT-MAE 2007-01)

2007-02 - GONÇALVES, M., Kolev, N. V., Fabris. A. E. Bounds for quantile-based measures of dependent risks' functions. 2007. 16p. (RT-MAE 2007-02)

2007-03 - MIMBELA, J. A. L, Kolev, N.V. Occupation measure of Markov - modulated risk processes 2007. 12p. (RT-MAE 2007-03)

2007-04 - GONÇALVES, M., Kolev, N. V., Fabris, A. E. Bounds for distorted risk measure. 2007. 11p. (RT-MAE 2007-04)

2007-05 - BUENO, V.C., A series representation of a parallel systems. 2007. 10p. (RT-MAE 2007-05)

2007-06 - KOLEV, N.V., PAIVA, D. Copula-Based Regression Models. 2007. 27p. (RT-MAE 2007-06)

2007-07 - POLETO, F. Z., SINGER, J.M., PAULINO, C.D., A product-multinomial framework for categorical data analysis with missing responses. 2007. 27p (RT-MAE 2007-07)

2007-08 - KOLEV, N.V., GONCALVES, M. DIMITROV, B. Probabilistic Properties of Sibuya's Dependence Function. 2007. 13p. (RT-MAE-2007-08)

2007-09 - AZEVEDO, C.L.N., ANDRADE, D.F. Latent Trait Estimation in Nominal Response Model: Latent Trait Asymmetry and Hierarchical and Empirical Framework. 2007.20p. (RT-MAE-2007-09)

2007-10 - AZEVEDO N.C.L., ANDRADE, D.F. Latent Trait Estimation Through Nominal Response and Dichotomous models: a case Study in a Multiple choice test.2007. 17p. (RT-MAE-2007-10)

2007-11- FIGUEIREDO, C.C., SANDOVAL, M.C., BOLFARINE,H., LIMA, C.R.O.P., Skew-Normal Linear Calibration: A Bayesian Perspective. 2007.18p.(RT-MAE-2007-10)

2007-12 - BAYES, C.L., BRANCO, M.D., Robust Linear Mixed Model using the Skew-t Distribution. 2007.18p. (RT-MAE-2007-11.)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

*Departamento de Estatística
IME-USP
Caixa Postal 66.281
05311-970 - São Paulo, Brasi*