Rev. Saúde públ., S. Paulo
2(1):1-9, jun. 1968

REGRESSÃO LINEAR COM DUPLO TRUNCAMENTO NA DISTRIBUIÇÃO DA VARIÁVEL DEPENDENTE [1]

José Maria Pacheco de SOUZA

Do problema do ajuste de uma regressão linear, quando a distribuição da variável dependente tem duplo truncamento, utilizando a função de máxima verossimilhança e um processo iterativo.

1 — INTRODUÇÃO

Motivados pela necessidade de estudar o comportamento da idade da mulher casada na época do primeiro abôrto em função da sua idade ao casar (MILANESI [3], 1968), ou seja, estabelecer a regressão entre estas variáveis, nos defrontamos com um problema envolvendo uma variável que, pela sua própria natureza, possui um campo de variação restrito a um sub-conjunto do total. Estamos nos referindo à variável — idade da mulher casada na época do primeiro abôrto — que tem seu campo de variação limitado, de um lado, pela idade ao casar, e de outro, pela idade na menopausa, ou seja, tem o que se denomina um duplo truncamento.

A solução do problema proposto se enquadra, portanto, no capítulo de regressão com duplo truncamento na distribuição da variável dependente. HOLGATE [2] (1965) já considerou o caso de regressão com truncamento simples; o presente trabalho representa uma extensão dos resultados daquele autor, para o caso mais geral de duplo truncamento.

2 — A FUNÇÃO DENSIDADE DE PROBABILIDADES

No que segue faremos abstração da motivação já apresentada, a fim de tratarmos do problema com maior generalidade.

Recebido para publicação em 16-12-1968.

(1) Da Cadeira de Estatística Aplicada à Saúde Pública da Faculdade de Higiene e Saúde Pública da USP.

Sejam x a variável independente e y a variável aleatória dependente distribuída normalmente, com média $a + bx$ e variância σ^2, ou, abreviadamente:

$$y \div N\ (a + bx;\ \sigma^2).$$

Havendo duplo truncamento, a função densidade de probabilidde de y, para cada x é:

$$f\ (y/x) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{y - a - bx}{\sigma}\right)^2} \cdot \frac{1}{I\ (w, z)}$$

onde

$$I\ (w, z) = \int_{\frac{w - a - bx}{\sigma}}^{\frac{z - a - bx}{\sigma}} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{t^2}{2}}\ dt,$$

$$w \leq y \leq z\ ;$$

ou seja, w e z são os pontos inferior e superior de truncamento na distribuição de y.

3 — ESTIMAÇÃO DOS PARÂMETROS a E b DA RETA DE REGRESSÃO E DA VARIÂNCIA σ^2

A estimação será feita pelo método da máxima verossimilhança. Para tanto, suponhamos que, em correspondência a cada x_i $(i = 1, 2, \quad m)$ fôsse tomada uma amostra de tamanho n_i, isto é, de valores de y_{ij} $(j = 1, 2, \quad ., n_i)$. Sejam:

$$\sum_{i=1}^{m} n_i = N, \qquad -\frac{a}{\sigma} = \xi, \qquad -\frac{b}{\sigma} = k.$$

A função de verossimilhança da amostra é:

$$L = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{y_{ij}}{\sigma} + \xi + kx_i\right)^2} \; \frac{1}{I(w_i, z_i)}$$

Tomando-se logaritmos naturais, temos:

$$\log L = L^* = -N \log \sigma - \frac{N}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\frac{y_{ij}}{\sigma} + \xi + kx_i\right)^2 - \sum_{i=1}^{m} n_i \log I(w_i, z_i)$$

Calculando-se as derivadas parciais para cada parâmetro, obtemos:

$$\frac{\partial L^*}{\partial \xi} = f(\xi) = -\frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}}{\sigma} - N\xi - k \sum_{i=1}^{m} n_i x_i + \sum_{i=1}^{m} n_i \frac{e(w_i) - e(z_i)}{I(w_i, z_i)} \qquad (1)$$

$$\frac{\partial L^*}{\partial k} = g(k) = -\frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij} x_i}{\sigma} - \xi \sum_{i=j}^{m} n_i x_i - k \sum_{i=1}^{m} n_i x_i^2 + \sum_{i=j}^{m} n_i x_i \frac{e(w_i) - e(z_i)}{I(w_i, z_i)} \qquad (2)$$

$$\frac{\partial L^*}{\partial \sigma} = h(\sigma) = -\frac{N}{\sigma} + \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}^2}{\sigma^3} + \frac{\xi}{\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}^2 + \frac{k}{\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_i y_{ij} + \frac{1}{\sigma^2} \sum_{i=1}^{m} n_i \frac{w_i \cdot e(w_i) + z_i \cdot e(z_i)}{I(w_i, z_i)} \qquad (3)$$

onde $e(w_i)$, $e(z_i)$ são as ordenadas obtidas na curva normal, nos pontos de truncamento w_i e z_i.

Derivando-se uma segunda vez, obtemos:

$$\frac{\partial f}{\partial \xi} = \sum_{i=1}^{m} n_i \frac{I_{(w_i,z_i)}[E_{(z_i)}\cdot e_{(z_i)} - E_{(w_i)}\cdot e_{(w_i)}] + [e_{(z_i)}]^2 + [e_{(w_i)}]^2 - 2[e_{(z_i)}\cdot e_{(w_i)}]}{[I_{(w_i,z_i)}]^2} - N \tag{4}$$

$$\frac{\partial f}{\partial k} = \frac{\partial g}{\partial \xi} = \sum_{i=1}^{m} n_i x_i \frac{I_{(w_i,z_i)}[E_{(z_i)}\cdot e_{(z_i)} - E_{(w_i)}\cdot e_{(w_i)}] + [e_{(z_i)}]^2 + [e_{(w_i)}]^2 - 2[e_{(z_i)}\cdot e_{(w_i)}]}{[I_{(w_i,z_i)}]^2} - \sum_{i=1}^{m} n_i x_i \tag{5}$$

$$\frac{\partial f}{\partial \sigma} = \frac{\partial h}{\partial \xi} = \frac{1}{\sigma^2} \sum_{i=1}^{m} n_i \frac{I_{(w_i,z_i)}[w_i\cdot E_{(w_i)}\cdot e_{(w_i)} - z_i\cdot E_{(z_i)}\cdot e_{(z_i)}] - z_i\cdot[e_{(z_i)}]^2 - w_i[e_{(w_i)}]^2 + e_{(z_i)}\cdot e_{(w_i)}[z_i + w_i]}{[I_{(w_i,z_i)}]^2} + \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}}{\sigma^2} \tag{6}$$

$$\frac{\partial g}{\partial k} = \sum_{i=1}^{m} n_i x_i^2 \frac{I_{(w_i,z_i)}[E_{(z_i)}(e_{(z_i)} - e_{(w_i)})] + [e_{(z_i)}]^2 + [e_{(w_i)}]^2 - 2[e_{(z_i)}\cdot e_{(w_i)}]}{[I_{(w_i,z_i)}]^2} - \sum_{i=1}^{m} n_i x_i^2 \tag{7}$$

$$\frac{\partial g}{\partial} = \frac{\partial h}{\partial k} = \frac{1}{\sigma^2} \sum_{i=1}^{m} n_i x_i \frac{I_{(w_i,z_i)}[w_i\cdot E_{(w_i)}\cdot e_{(w_i)} - z_i\cdot E_{(z_i)}\cdot e_{(z_i)}] - z_i\cdot[e_{(z_i)}]^2 - w_i[e_{(w_i)}]^2 + e_{(z_i)}\cdot e_{(w_i)}[z_i + w_i]}{[I_{(w_i,z_i)}]^2} + \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij} x_i}{\sigma^2} \tag{8}$$

$$\frac{\partial h}{\partial \sigma} = \frac{N}{\sigma^2} - \frac{2}{\sigma^3} \sum_{i=1}^{m} n_i \frac{z_i \cdot e_{(z_i)}}{I_{(w_i,z_i)}} + \frac{1}{\sigma^4} \sum_{i=1}^{m} n_i \cdot z_i \frac{I_{(w_i,z_i)}\left[E_{(z_i)} \cdot e_{(z_i)} \cdot z_i\right] + \left[e_{(z_i)}\right]^2 \cdot z_i - w_i \cdot e_{(z_i)} \cdot e_{(w_i)}}{\left[I_{(w_i,z_i)}\right]^2} -$$

$$- \frac{2}{\sigma^3} \sum_{i=1}^{m} n_i \frac{w_i \quad e_{(w_i)}}{I_{(w_i,z_i)}} + \frac{1}{\sigma^4} \sum_{i=1}^{m} n_i w_i \frac{I_{(w_i,z_i)} \left[E_{(w_i)} \cdot e_{(w_i)} \cdot w_i\right] - \left[e_{(w_i)}\right]^2 \cdot w_i + z_i \cdot e_{(z_i)} \cdot e_{(w_i)}}{\left[I_{(w_i,z_i)}\right]^2} -$$

$$- \frac{1}{\sigma^4} \left[3 \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}^2 + 2\sigma\xi \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij} + 2\sigma k \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij} x_i \right] \qquad (9)$$

onde:

$$E_{(\ w_i\ )} = \frac{(w_i)}{\sigma} + \xi + kx_i$$

$$E_{(\ z_i\ )} = \frac{(z_i)}{\sigma} + \xi + kx_i .$$

Em virtude da dificuldade de resolver estas equações de forma explícita, os estimadores $\overset{\Lambda}{a}$, $\overset{\Lambda}{b}$, $\overset{\Lambda}{\sigma}$, serão obtidos a partir dos resultados numéricos das equações 1 a 9, por processo iterativo, utilizando-se então o método de Newton-Raphson (WHITTAKER & ROBINSON [7], 1926).

Como primeiro passo, calculam-se valores $\overset{\Lambda}{\xi_1}$, $\overset{\Lambda}{k_1}$ e $\overset{\Lambda}{\sigma_1}$, pelo processo clássico de regressão e análise de variância, e, com êstes valores, resolve-se o seguinte sistema de 3 equações a 3 incógnitas:

$$\begin{aligned} f(\xi) + p\frac{\partial f}{\partial \xi} + q\frac{\partial f}{\partial k} + r\frac{\partial f}{\partial \sigma} &= 0 \\ g(k) + p\frac{\partial g}{\partial \xi} + q\frac{\partial g}{\partial k} + r\frac{\partial g}{\partial \sigma} &= 0 \\ h(\sigma) + p\frac{\partial h}{\partial \xi} + q\frac{\partial h}{\partial k} + r\frac{\partial h}{\partial \sigma} &= 0 , \end{aligned} \qquad (10)$$

onde as derivadas são tomadas nos pontos $\xi = \overset{\Lambda}{\xi_1}$; $k = \overset{\Lambda}{k_1}$ e $\sigma = \overset{\Lambda}{\sigma_1}$, obtendo-se valores de p_1, q_1, r_1, respectivamente, os acréscimos de $\overset{\Lambda}{\xi_1}$, $\overset{\Lambda}{k_1}$ e $\overset{\Lambda}{\sigma_1}$ na 1.ª iteração. Assim, passamos a ter os seguintes novos estimadores de ξ, k e σ:

$$\overset{\Lambda}{\xi_2} = \overset{\Lambda}{\xi_1} + p_1 \qquad (11)$$

$$\overset{\Lambda}{k_2} = \overset{\Lambda}{k_1} + q_1 \qquad (12)$$

$$\overset{\Lambda}{\sigma_2} = \overset{\Lambda}{\sigma_1} + r_1 \qquad (13)$$

Os resultados de 11, 12, 13 é que serão utilizados agora num 2.º ciclo interativo para solução do sistema 10 (onde as derivadas serão tomadas nos pontos $\xi = \overset{\Lambda}{\xi_2}$; $k = \overset{\Lambda}{k_2}$; $\sigma = \overset{\Lambda}{\sigma_2}$) e obtenção de novos acréscimos e, portanto, novos estimadores. O processo se encerra quando os acréscimos (p_n, q_n, r_n) se tornarem desprezíveis, e então:

$$\overset{\Lambda}{\xi}_{n+1} = \overset{\Lambda}{\xi_n}; \quad \overset{\Lambda}{k}_{n+1} = \overset{\Lambda}{k_n}; \quad \overset{\Lambda}{\sigma}_{n+1} = \overset{\Lambda}{\sigma_n}$$

Nestas condições, os estimadores de a, b e σ serão:

$$\hat{a} = -\hat{\xi}_n \hat{\sigma}_n$$
$$\hat{b} = -\hat{k}_n \hat{\sigma}_n$$
$$\hat{\sigma} = \sigma_n, \qquad \text{e a reta de regressão:}$$
$$\hat{Y}_i = \hat{a} + \hat{b}\, x_i.$$

4 — CONSIDERAÇÕES SÔBRE OS RESULTADOS

Nossos resultados (equações 1 a 9) diferem dos de HOLGATE (equações 3 e 4) devido ao têrmo $\nu(x)$, que é o quociente da ordenada no ponto de truncamento pela área à sua direita; no caso em aprêço temos truncamento, e êste fato nos conduz a um quociente em que temos, no numerador, a diferença entre as ordenadas dos pontos de truncamento, e no denominador, a área entre os 2 pontos de truncamento.

Como conseqüência, temos também as divergências devidas ao têrmo $\lambda(x) = \nu'(x)$, para nós substituído por equações mais complexas.

Essas diferenças não são de todo eliminadas, mesmo quando um dos pontos de truncamento está muito afastado da média, determinando uma ordenada pràticamente igual a zero.

Assim sendo, ainda nestes casos mais favoráveis, não podemos nos utilizar das tabelas devidas a SAMPFORD [6] (1952), que nos dão valores de $\nu(x)$ e $\lambda(x)$, e devemos recorrer às tabelas da curva normal para áreas e ordenadas (PEARSON & HARTLEY [5], 1958 e BOLL [1], 1947).

5 — EXEMPLO

Retomemos o problema que motivou a generalização aqui apresentada, isto é, estabelecer a reta de regressão entre a "idade da mulher na época do primeiro abôrto" (y) e a "idade da mulher ao casar" (x); MILANESI [3] (1968), obteve, pelo método dos mínimos quadrados, $\hat{Y}_1 = 12{,}46 + 0{,}696\ x$; $\hat{\sigma}_1 = 4{,}90$, observando 306 mulheres casadas que tinham tido abôrto.

Como a distribuição da primeira variável pode ser considerada duplamente truncada, foi utilizado, a seguir, o método aqui descrito para a devida correção, encontrando-se os seguintes valores para as equações 1 a 9:

$$f(\xi) = 47{,}556$$
$$g(k) = 1.047{,}251$$
$$h(\sigma) = 97{,}262$$
$$\frac{\partial f}{\partial \xi} = -231{,}653$$
$$\frac{\partial f}{\partial k} = -4.690{,}929$$

$$\frac{\partial f}{\partial \sigma} = 279{,}302$$

$$\frac{\partial g}{\partial k} = -\ 236.276{,}135$$

$$\frac{\partial g}{\partial \sigma} = 5.844{,}054$$

$$\frac{\partial h}{\partial \sigma} = -\ 484{,}819,$$

com a nova equação sendo:

$$\hat{Y}_2 = 12{,}31 + 0{,}655\ x; \quad \hat{\sigma}_2 = 4{,}61.$$

Êstes resultados são muito semelhantes aos anteriores, indicando que não se fazia necessário um nôvo ciclo iterativo.

Usando o método de estimação da máxima verossimilhança, a matriz de variância-covariância dos estimadores de ξ, k e σ será dada por (MOOD & GRAYBILL [4], 1963):

$$\Delta_{(\hat{\xi},\ \hat{k},\ \hat{\sigma})} = \begin{bmatrix} -\frac{\partial f}{\partial \xi} & -\frac{\partial f}{\partial k} & -\frac{\partial f}{\partial \sigma} \\ -\frac{\partial f}{\partial k} & -\frac{\partial g}{\partial k} & -\frac{\partial g}{\partial \sigma} \\ -\frac{\partial f}{\partial \sigma} & -\frac{\partial g}{\partial \sigma} & -\frac{\partial h}{\partial \sigma} \end{bmatrix}^{-1}$$

Denotando por $D_{(\hat{a},\ \hat{b},\ \hat{\sigma})}$ a matriz de variância-covariância dos estimadores de a, b e σ, e por J a matriz de transformação:

$$J = \begin{bmatrix} -\hat{\sigma} & 0 & -\hat{\xi} \\ 0 & -\hat{\sigma} & -\hat{k} \\ 0 & 0 & 1 \end{bmatrix}$$

tem-se:

$$D_{(\hat{a},\ \hat{b},\ \hat{\sigma})} = J\ \Delta\ J',$$

que no exemplo considerado assumem os seguintes valores:

$$D_{(\hat{a},\,\hat{b},\,\hat{\sigma})} = \begin{bmatrix} -4{,}9 & 0 & 2{,}543 \\ 0 & -4{,}9 & 0{,}142 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0{,}01662 & -0{,}00013 & 0{,}00797 \\ -0{,}00013 & 0{,}00001 & 0{,}00001 \\ 0{,}00797 & 0{,}00001 & 0{,}00676 \end{bmatrix} =$$

$$= \begin{bmatrix} -4{,}9 & 0 & 0 \\ 0 & -4{,}9 & 0 \\ 2{,}543 & 0{,}142 & 1 \end{bmatrix} = \begin{bmatrix} 0{,}24403 & -0{,}00640 & -0{,}02187 \\ -0{,}00640 & 0{,}00029 & 0{,}00092 \\ -0{,}02187 & 0{,}00092 & 0{,}006757 \end{bmatrix},$$

onde vemos que a variância de $\hat{a} = 0{,}24403$ e a de $\hat{b} = 0{,}00029$, enquanto que a de $\hat{\sigma} = 0{,}00657$.

SUMMARY

A solution to the problem of fitting a linear regression with double truncation in the distribution of the dependent variable is obtained, using the maximum likelihood function and an iterative process.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BOLL, M. — *Tables numériques universelles: des laboratoires et bureaux d'etude.* Paris, Dunod, 1947.

2. HOLGATE, P. — Fitting a straight line to data from a truncated population. *Biometrics* 21(3):715-720, Sept. 1965.

3. MILANESI, M. L. — *Abôrto provocado.* São Paulo, 1968. (Tese de doutoramento — Fac. Hig. Saúde Públ. Univ. S. Paulo).

4. MOOD, A. M. & GRAYBILL, F. A. — *Introduction to the theory of statistics.* 2nd ed. New York, Mc-Graw-Hill, 1963. p. 236-237.

5. PEARSON, E. S. & HARTLEY, H. O., ed. — *Biometrika tables for statisticians.* 2nd ed. Cambridge, University Press, 1958. v. 1.

6. SAMPFORD, M. R. — The estimation of response-time distributions. II. Multistimulus distributions. *Biometrics,* 8(4):307-369, Dec. 1952.

7. WHITTAKER, F. T. & ROBINSON, G. — *The calculus of observations: a treatese on numerical mathematics.* 2nd ed. London, Blackie, 1926.