Estimação Bayesiana em Modelos de Regressão Complemento Log-Log

Maria Regina Madruga, **Pedro Silvestre da Silva Campos**,
Faculdade de Estatística, ICEN, UFPA,
66600-000, Belém, PA
E-mail: madruga@ufpa.br, psscam@yahoo.com.br,

Resumo: *Este trabalho utiliza os métodos de estimação Bayesiana em Modelos de Regressão Complemento Log-Log, também conhecidos como regressão* extremito. *O processo de estimação baseia-se nos trabalhos de [1] e [5], que utilizam variáveis latentes no processo de estimação dos parâmetros dos modelos de Regressão Probito e Logístico, respectivamente, a partir do Amostrador de Gibbs([3]). São feitas aplicações em dados categorizados já utilizados na literatura em ajustes com outros modelos, para comparação das metodologias.*

1 Introdução

A metodologia de estimação desenvolvida neste trabalho está baseada nos trabalhos de [1] e [5], que fazem uso de variáveis latentes no processo de estimação dos parâmetros em modelos de regressão com respostas categóricas.
Na proposta de [1] ajusta-se o modelo probit binário, com a introdução de uma sequência de n variáveis latentes, $z_1, z_2, \cdots, z_n$, com $z_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1)$, em que $\boldsymbol{x}_i^T$ é o vetor de covariáveis para a i-ésima observação e $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo, e define-se a variável resposta binária para a i-ésima observação como

$$Y_i = \begin{cases} 0, & \text{se } z_i \leq 0 \\ 1, & \text{se } z_i > 0 \end{cases} .$$

Como as variáveis z_i, $i = 1, \cdots, n$, não são observadas, eles mostram que dado o valor da variável oservável Y_i, tem-se que $z_i|\beta, \sigma^2$ tem distribuiçao normal truncada. A técnica sugerida em [5] está baseada em uma proposta de aproximação da distribuição a posteriori obtida em [1], usando o método *data augumentation*, uma função de ligação *logit* e variáveis latentes com distribuição uniforme, com implementação via Amostrador de Gibbs ([3]; [2]) para o processo de simulação e estimação dos parâmetros.

2 Amostrador de Gibbs

O Amostrador de Gibbs é um método de amostragem iterativo de uma cadeia de Markov, cuja transição de um estado a outro é feita a partir das distribuições condicionais completas posteriores de um vetor de parâmetros $\boldsymbol{\theta}$ de dimensão $n \times 1$. Define-se a distribuição condicional completa de um componente qualquer θ_i como a distribuição condicional deste, dado todos os outros parâmetros e os dados, denotada por $p_i\left(\theta_i|\boldsymbol{\theta}_{-\theta_i}, Y\right)$, com $\boldsymbol{\theta}_{-\theta_i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_n)$. A atualização feita pelo amostrador de Gibbs é um caso particular do algoritmo de Metropolis-Hastings ([8]).

A implentação do Amostrador de Gibbs, na iteração k, deve obedecer aos seguintes passos:
Passo 1: $\theta_1^{(k)} \sim p_1\left(\theta_1|\theta_2^{(k-1)}, \theta_3^{(k-1)}, ..., \theta_n^{(k-1)}, Y\right)$
Passo 2: $\theta_2^{(k)} \sim p_2\left(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, ..., \theta_n^{(k-1)}, Y\right)$
Passo 3: $\theta_3^{(k)} \sim p_3\left(\theta_3|\theta_1^{(k)}, \theta_2^{(k)}, ..., \theta_n^{(k-1)}, Y\right)$
$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$
Passo n: $\theta_n^{(k)} \sim p_n\left(\theta_n|\theta_1^{(k)}, \theta_2^{(k)}, ..., \theta_{n-1}^{(k)}, Y\right)$.
Repita os passos $1, 2, 3, ..., n$ para $k = 1, 2, 3, ...$

3 Resposta Dicotômica

Seja y_i uma variável aleatória dicotômica, tal que

$$y_i = \begin{cases} 1, & \text{com probabilidade } \pi_i; \\ 0, & \text{com probabilidade } 1 - \pi_i. \end{cases}$$

Segundo [7] e [9] tem-se que o modelo de *Regressão Extremito* (**RE**) faz uso da função de ligação *complemento log-log*, ou seja, $\log[-\log(1-\pi_i)]$, para modelar π_i associada a um vetor de p covariáveis $\mathbf{X} = (\mathbf{1}, X_{i1}, \cdots, X_{ip})$. Assim, o modelo é dado por

$$\pi_i = 1 - e^{-e^{\beta \mathbf{X}}} \tag{1}$$

onde $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$.

Tem-se que **RE** dado em (1) é a função de distribuição acumulada da distribuição Gumbel para valores extremos, isto é,

$$\pi_i = F(z) = \int_{-\infty}^{\beta \mathbf{X}} e^z e^{-e^z} dz, -\infty < z < \infty. \tag{2}$$

De forma análoga a [5], introduzimos no modelo uma variável latente independente $U \sim Uniforme(0,1)$. Neste caso, segue que

$$\begin{aligned} \pi_i = F(z) &= \int_{-\infty}^{\boldsymbol{\beta X}} e^z e^{-e^z} dz & (3) \\ &= P\left(U < 1 - e^{-e^{\beta \mathbf{X}}}\right). & (4) \end{aligned}$$

A variável latente introduzida no modelo será gerada para cada observação, dando origem ao vetor de variáveis latentes $\mathbf{u}= (u_1, \cdots, u_n)$. A densidade conjunta do vetor de parâmetros $\boldsymbol{\beta}$ e $\mathbf{u}$, dado o vetor de observações $\mathbf{Y} = (\mathbf{y_1}, \cdots, \mathbf{y_n})$, será dada por

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{Y}) &\propto \pi(\boldsymbol{\beta}) L(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{Y}) \\ &\propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Bigg[I\left(u_i \leq 1 - e^{-e^{\boldsymbol{\beta X}}}\right) \\ &\quad I(y_i = 1) \\ &\quad + I\left(u_i > 1 - e^{-e^{\boldsymbol{\beta X}}}\right) I(y_i = 0) \Bigg] \\ &\quad I(0 \leq u_i \leq 1) \end{aligned}$$

onde $\pi(\boldsymbol{\beta}) \propto 1$ é a priori do vetor de parâmetros e $I(X \in A)$ é a função indicadora, que é igual a 1 se $X \in A$ e 0, caso contrário, sendo assim;

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{Y}) &\propto \prod_{i=1}^{n} \left[I\left(\boldsymbol{\beta X} \geq \log\left[-\log(1-u_i)\right]\right) \right. \\ &\quad I(y_i = 1) \\ &\quad \left. + I\left(u_i > 1 - e^{-e^{\boldsymbol{\beta X}}}\right) I(y_i = 0) \right] \\ &\quad I(0 \leq u_i \leq 1). \end{aligned} \tag{5}$$

De (5), se $y_i = 1$ segue que:

$$\beta_t \geq \frac{1}{x_{it}} \left\{ \log\left[-\log(1-u_i)\right] - \sum \beta_k x_{ik} \right\}$$

para todo i com $y_i = 1$ e $x_{it} > 0$, sendo válida também se $y_i = 0$ e $x_{it} < 0$.

Da mesma forma, se $y_i = 0$, segue que:

$$\beta_t \leq \frac{1}{x_{it}} \left\{ \log\left[-\log(1-u_i)\right] - \sum \beta_k x_{ik} \right\}$$

para todo i com $y_i = 0$ e $x_{it} > 0$, sendo válida também se $y_i = 1$ e $x_{it} < 0$.

Sendo assim, observa-se o surgimento natural de dois conjuntos indicadores, A_t e B_t, que auxiliam na implementação do Amostrador de Gibbs. Os conjuntos A_t e B_t são definidos por

$$\begin{aligned} A_t &= \{i : [(y_i = 1) \cap (x_{it} > 0)] \cup \\ &\quad [(y_i = 0) \cap (x_{it} < 0)]\} \end{aligned}$$

e

$$\begin{aligned} B_t &= \{i : [(y_i = 0) \cap (x_{it} > 0)] \cup \\ &\quad [(y_i = 1) \cap (x_{it} < 0)]\} \end{aligned}$$

Considerando que não há nenhum conhecimento prévio sobre $\boldsymbol{\beta}$, assume-se uma priori difusa para $\boldsymbol{\beta}$, isto é, $\pi(\boldsymbol{\beta}) \propto 1$. Tem-se, então, que a distribuição completa de β_t será uma distribuição uniforme

$$\beta_t | \boldsymbol{\beta}_{(-t)}, u, Y \sim Uniforme(a_k, b_k), t = 0, 1, \cdots, p.$$

sendo

$$a_t = \max_{i \in A_t} \left\{ \frac{1}{x_{it}} \left\{ \log\left[-\log(1-u_i)\right] - \sum \beta_k x_{ik} \right\} \right\}$$

e

$$b_t = \min_{i \in B_t} \left\{ \frac{1}{x_{it}} \left\{ \log\left[-\log(1-u_i)\right] - \sum \beta_k x_{ik} \right\} \right\}.$$

4 Resposta Multinomial Ordinal

Seja y_i uma variável aleatória multinomial assumindo valores em r categorias ordenadas, tal que $P(y_i = j) = \pi_{ij}$, $j = 1, 2, \cdots, r$, e a probabilidade acumulada até a $j-$ésima categoria denotada por

$$\eta_{ij} = \sum_{k=1}^{j} \pi_{ik} = P(y_i \leq j).$$

Segundo [9] tem-se que para modelos com respostas ordinais o modelo extremito é dado por

$$\eta_{ij} = 1 - e^{-e^{\alpha_j + \boldsymbol{\beta X}}} \tag{6}$$

onde $\boldsymbol{X} = (1, X_{i1}, \cdots, X_{ip})$ é o vetor de covariadas e $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ é o vetor de coeficientes de $\boldsymbol{X}$ e $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_r)$ é o vetor de pontos de corte, tal que $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_r = \infty$.

A distribuiçào conjunta de $\boldsymbol{\alpha}, \boldsymbol{\beta}$ e $\boldsymbol{u}$, dado $\boldsymbol{Y}$, é dada por

$$\begin{aligned} \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{u} | \boldsymbol{Y}) \quad \propto \quad & \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \prod_{i=1}^{n} \left\{ \sum_{j=1}^{r-1} I(y_i = j) \right. \\ & \left. I(\eta_{i,j-1} < u_i \leq \eta_{i,j}) \right\} \\ & I(0 \leq u_i \leq 1). \end{aligned} \tag{7}$$

Assumindo uma priori conjunta difusa para $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$, tem-se que a distribuição condicional de u_i, dado $\boldsymbol{\alpha}, \boldsymbol{\beta}$ e $y_i = j$, é dada por

$$u_i | \alpha, \beta, y_i = j \sim Uniforme(\eta_{i,j-1}, \eta_{i,j}) \tag{8}$$

com $i = 1, \cdots, n$ e η_{ij} dada por (6).

Introduzindo em (6) a variável latente com distribuição uniforme no intervalo [0,1], de forma análoga à Seção 3, tem-se que

$$\eta_{ij} = P\left(U_i < 1 - e^{-e^{\alpha_j + \boldsymbol{\beta X}}} \right),$$

de onde segue, para os parâmetros das covariadas, que

$$\beta_t < \frac{1}{x_{it}} \left\{ \log[-\log(1 - u_i)] - \alpha_{j-1} - \sum_{k \neq t}^{p} \beta_k x_{ik} \right\}$$

e

$$\beta_t > \frac{1}{x_{it}} \left\{ \log[-\log(1 - u_i)] - \alpha_j - \sum_{k \neq t}^{p} \beta_k x_{ik} \right\}.$$

De forma a sumarizar a notação, denota-se por

$$T_{ij} = \frac{1}{x_{it}} \left\{ \log[-\log(1 - u_i)] - \alpha_j - \sum_{k \neq t}^{p} \beta_k x_{ik} \right\},$$

logo, tem-se que $T_{i,j,t} < \beta_t < T_{i,j-1,t}$ para todo $y_i = j$ e $x_{it} > 0$, e $T_{i,j-1,t} < \beta_t < T_{i,j,t}$ para todo $y_i \neq j$ e $x_{it} < 0$, surgindo assim naturalmente o conjunto $A_j = \{i : y_i = j\}$.

Sendo a distribuição completa e truncada,de β_t, dada por

$$\beta_t | \boldsymbol{\beta}_{(-t)}, \boldsymbol{\alpha}, \boldsymbol{u}, \boldsymbol{y} \sim Uniforme(a_t, b_t)$$

com $t = 1, \cdots, p$, tal que

$$a_t = \max_j \left\{ \max_{i \in A_j} [\min(T_{i,j-1,t}; T_{i,j,t})] \right\}$$

e

$$b_t = \min_j \left\{ \min_{i \in A_j} [\max(T_{i,j-1,t}; T_{i,j,t})] \right\}.$$

Na determinação da distribuição condicional dos pontos de corte, $\alpha' s$, deve-se observar as condições $u_i \leq \eta_{ij}$ para todo $i \in A_j$ e $u_i > \eta_{ij}$ para todo $i \in A_{j+1}$ e, também, que $\alpha_{j-1} < \alpha_j < \alpha_{j+1}$. Sendo assim, segue que

$$\alpha_j | \alpha_{(-j)}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y} \sim Uniforme(c_j, d_j)$$

onde

$$c_j = \max_{i \in A_{j+1}} \{\max[\log[-\log(1 - u_i)] - \boldsymbol{\beta X}; \alpha_{j-1}]\}$$

e

$$d_j = \min_{i \in A_j} \{\min[\log[-\log(1 - u_i)] - \boldsymbol{\beta X}; \alpha_{j+1}]\}$$

5 Aplicação 1

Um conjunto de dados, conhecido na literatura de modelos dose-resposta, encontra-se em [2], e baseia-se no comportamento de besouros adultos face à exposição a dissulfeto de carbono (CS_2) durante 5 horas. A curva de dose-resposta da mortalidade dos besouros foi formada a partir de 8 dosagens, e foi ajustada segundo a metodologia da Seção 3, devido ao fato dos dados sugerirem um comportamento anômalo em uma das extremidades. Os dados encontram-se na Tabela 1, onde as três colunas correspondem, respectivamente, ao número de besouros observados (n_i), ao número de besouros mortos r_i e ao log de cada dosagem de CS_2, $i = 1, 2, ..., 8$.

Tabela 1: Dose-resposta

n_i	r_i	$\log(Dose_i)$
59	6	1,6907
60	13	1,7242
62	18	1,7552
56	28	1,7842
63	52	1,8113
59	53	1,8369
62	61	1,8610
60	60	1,8839

A Figura 1 mostra o gráfico do modelo dose-resposta ajustado para os dados da Tabela 1, segundo a metodologia da Seção 3, dado por

$$\pi_i = 1 - e^{-e^{-39,5475 + 22,0273X}}.$$

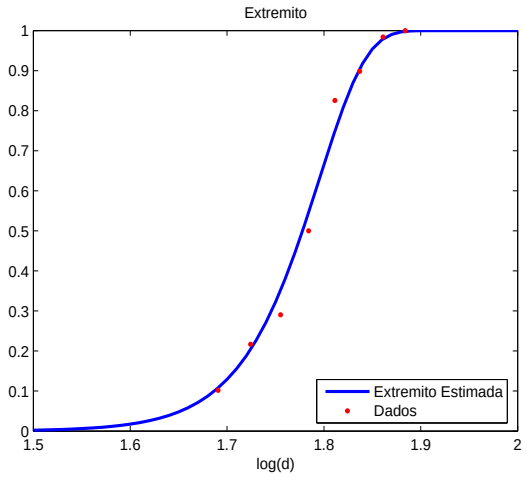

Figura 1: Função Extremito ajustada

Estes dados, quando ajustados pela metodologia apresentada em [5], fez uso de uma transformação exponencial da dose, isto é, $t_i = \exp(x_i)$, e o modelo estimado é dado por

$$\pi_i = \frac{\exp(-34,3086 + 5,8586t_i)}{1 + \exp(-34,3086 + 5,8586t_i)},$$

que é apresentado na Figura 2. Nota-se que o

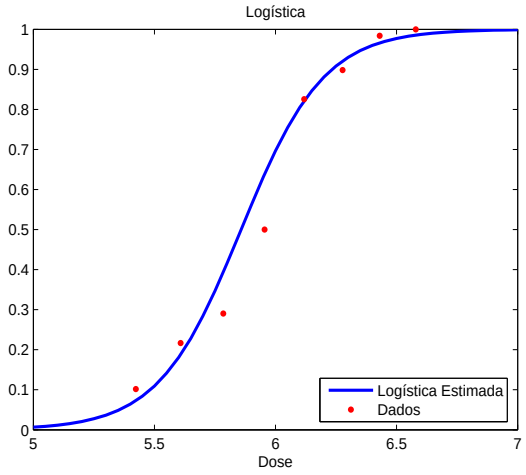

Figura 2: Proporção de Besouros mortos expostos à (CS_2)

ajuste proposto na Seção 3, modelo de Regressão Extremito (Figura 1), apresenta um melhor desempenho do que o Modelo Logístico ajustado (Figura 2), com técnicas de estimação Bayesiana semelhantes, fazendo uso de variáveis latentes no processo de implementação do Amostrador de Gibbs.

6 Aplicação 2

[6] propôs um modelo de Regressão Logística Multinomial para dados de dose-resposta de um experimento em dosimetria citogenética. O modelo de regressão logística proposto se caracteriza por um modelo linear inverso para a transformação *log-odds* da freqüência de aberrações, ou seja, a presença de micronúcleos(MN). Sendo π_{ij} a proporção de células com j $(j = 0, 1, 2)$ MN na i-ésima dose $(i = 1, 2, \cdots, 10)$, o modelo proposto é dado por:

$$\pi_{ij} = \frac{exp(H_j)}{1 + exp(H_1) + exp(H_2)}$$

tal que

$$H_j = -\left(\beta_{0j} + \frac{\beta_{1j}}{\beta_{2j} + D_i}\right), j = 1, 2.$$

Os dados $\boldsymbol{y}_i = (y_{i0}, y_{i1}, y_{i2})$ são apresentados na Tabela 2, com y_{ij} representando a frequência de células com j $(j = 0, 1, 2)$ MN na i-ésima dose $(i = 1, 2, \cdots, 10)$, e n_i o número total de células observadas na i-ésima dose.

Tabela 2: Dose-resposta

i	*Dose* (cGy)	y_{i0}	y_{i1}	y_{i2}	n_i
1	5	481	17	2	500
2	10	477	19	4	500
3	25	471	24	5	500
4	50	450	44	6	500
5	100	431	59	10	500
6	200	339	140	21	500
7	300	304	132	64	500
8	400	240	189	72	501
9	500	174	197	129	500
10	600	122	173	211	506

A implementação da Seção 4 está sendo feita para os dados de [6] e os resultados se mostram bastantes promissores, estando em fase de ajustes finais de comparação e ajuste dos parâmetros para o modelo de Regressão Extremito Multinomial Ordinal. Sendo necessário encontrar a melhor transformação da covariada para não colocar em dúvida a consistência e validade dos resultados.

Referências

[1] Albert, J. H; Chib, S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88, (1993) 669-679.

[2] Bliss, C. I. The calculation of the dosage-mortality curve. Annals Applied Biology, 22, (1935) 134-167.

[3] Gelfand, A. E.; Smith, A. F. M. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, (1990) 398-409.

[4] Geman, S.; Geman, D. Stochastic Relaxation, Gibbs Distribuitions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, (1984) 721-741.

[5] Groenewald, P. C. N; Molkgatlhe, L. Bayesian computation for logistic regression. *Computaional Statistics & Data Analysis*, 48, (2005) 857-868.

[6] Madruga, M. R.; Pereira, C. A. de B.; Gay-Rabelo, M. N. Bayesian dosimetry: radiation dose versus frequencies of cells with aberrations. *Envirometrics*, 5, (1994) 47-56.

[7] McCullagh, P.; Nelder, J. A. "Generalized Linear Models". Chapman and Hall, 2nd ed., London, 1989.

[8] Paulino, C. D.; Turkman, M. A. A.; Murteira, B. "Estatística Bayesiana". Fundação Calouste Gulbenkian, Lisboa, 2003.

[9] Paulino, C. D.; Singer, J. M. "Análise de Dados Categorizados". Edgard Blucher, São Paulo, 2006.