1 Método *bootstrap* aplicado à modelos lineares generalizados em experimentação agronômica

Rubem Kaipper Ceratti[1,2]
Afrânio Márcio Corrêa Vieira[1]
Joseane Padilha da Silva[2]
Cássio Costa da Silva Curi[2]

[1]Departamento de Estatística, Universidade de Brasília, Brasília, DF
[2]EMBRAPA Recursos Genéticos e Biotecnologia, Brasília, DF

Introdução

Nas últimas décadas, o desenvolvimento dos computadores permitiu que novas técnicas e métodos estatísticos fossem desenvolvidos para resolver problemas que antes eram extremamente complexos de serem resolvidos analiticamente pela Estatística Clássica. Problemas inferenciais gerados por amostras pequenas, como viés nas estimativas de parâmetros e distribuições sem aproximações assintóticas, passaram a ser abordados por métodos de reamostragem.

Em um cenário em que se trabalha com uma quantidade muito pequena de observações, frequentemente casos em que possuem uma dificuldade inerente para serem coletados - tempo, dinheiro, recursos humanos, etc -, a obtenção de resultados válidos e acurados a partir da pouca quantidade de informação se torna mais difícil. Embora esse problema possa ser tratado de forma analítica, a flexibilidade da abordagem dos métodos computacionalmente intensivos permite que se atinja resultados semelhantes, com a vantagem da simplicidade na aplicação do método.

Um exemplo deste tipo de situação é associado aos experimentos agronômicos, em que o tempo e custo envolvidos na obtenção das observações inviabilizam a coleta de muitos dados. E apesar dessa limitação, deseja-se utilizar a informação disponível para investigar fatores influentes na pesquisa, como na situação de comparação de diferentes tratamentos.

Metodologia

Família Exponencial e Modelos lineares generalizados

Seja uma amostra aleatória de tamanho n de pares de observações (x_i, y_i), compondo uma matriz $\mathbf{X}$ de dimensão n por *(p+1)* com p variáveis explicativas e Y um vetor de observações da variável resposta, $Y = (Y_1, Y_2, ..., Y_n)$. Assume-se que os Y_i são independentes e cada um tem densidade dada por

$$f(y_i; \theta_i, \phi) = exp\Big[\frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(Y_i, \phi)\Big] \quad (1)$$

com

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{2}$$

e

$$Var(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i) \tag{3}$$

sendo, pela notação de Smyth (1989), θ_i o parâmetro canônico, $a_i(\phi) = \phi/w_i$, ϕ o parâmetro de dispersão, w_i um peso "a priori"e $V(\mu_i)$ a função de variância dada por $V(\mu_i) = d\mu_i/d\theta_i$.

Os fatores e covariáveis organizados em X estão expressos no preditor linear na forma

$$\eta_i = x_i^T \beta \tag{4}$$

e a relação funcional entre a média da variável resposta Y e preditor linear é dada por

$$g(\mu_i) = \eta_i \tag{5}$$

onde β é o vetor dos parâmetros e $g(\cdot)$ é uma função monotônica e diferenciável, denominada função de ligação.

***Bootstrap* para Modelos Lineares Generalizados**

Davison e Hinkley (1997) descrevem uma abordagem semiparamétrica da aproximação Monte Carlo para os estimadores do modelo linear generalizado. Entende-se por abordagem semiparamétrica a utilização dos resíduos do modelo como objeto da reamostragem a ser realizada em cada ciclo do processo. Isto é, fixando-se a matriz $\mathbf{X}$ de variáveis explicativas e as estimativas $\hat{\mu}_i$, é possível reconstruir a variável resposta por meio da reamostragem dos resíduos.

Para modelos lineares generalizados, três resíduos podem ser utilizados: o resíduo de Pearson padronizado, o resíduo padronizado no preditor linear e o resíduo da componente de *deviance*.

Tomando o resíduo de Pearson padronizado, dado por

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{[\hat{\phi}V(\hat{\mu}_i)(1-h_i)]^{1/2}}, \qquad i = 1, ..., n \ , \tag{6}$$

sendo $\hat{\phi}$ o parâmetro de dispersão estimado, e h_i é o i-ésimo elemento diagonal da matriz H, definida como $H = X(X^TWX)^{-1}X^TW$.

Define-se então

$$y_i^* = \hat{\mu}_i + [\hat{\phi}V(\hat{\mu}_i)]^{1/2}\epsilon_i^* \tag{7}$$

em que $\epsilon_i^* = r_{P_i}^* - \bar{r}_P$ são os resíduos reamostrados.

Alternativamente, define-se o esquema para o resíduo paronizado no preditor linear:

$$r_{L_i} = \frac{g(y_i) - g(\hat{\mu}_i)}{[\hat{\phi}\dot{g}^2(\hat{\mu}_i)V(\hat{\mu}_i)(1-h_i)]^{1/2}}, \qquad i = 1, ..., n \ , \tag{8}$$

em que $\dot{g}(\mu)$ é a derivada primeira da função de ligação com relação à μ, ou seja, $\dot{g}(\mu) = d\eta/d\mu$. Dessa forma o vetor resposta é reconstruído fazendo-se

$$y_i^* = g^{-1}[x_i^T\hat{\beta} + \dot{g}(\hat{\mu}_i)[\hat{\phi}V(\hat{\mu}_i)]^{1/2}\epsilon_i^*] \tag{9}$$

em que $\epsilon_i^* = r_{L_i}^*$. Aqui, os resíduos não são ajustados pela média, a menos que a função de ligação seja a identidade.

A terceira abordagem, que se apresenta como alternativa às duas primeiras, é baseada na reamostragem dos resíduos da componente de deviance padronizados, que é escrito como

$$r_{D_i} = \frac{d_i}{(1-h_i)^{1/2}}, \qquad i = 1, ..., n \ , \tag{10}$$

sendo $d_i = d(y_i, \hat{\mu}_i)$ o resíduo da componente de *deviance*. Amostrando-se aleatoriamente os resíduos padronizados, o novo y_i^* é o valor que soluciona

$$\epsilon_i^* = d(y_i^*, \hat{\mu}_i)$$

onde $\epsilon_i^* = r_{D_i}^*$.

Experimento de conservação de sementes

Como aplicação da metodologia *bootstrap*, tem-se um conjunto de dados de tratamento de sementes de *Piper aduncum*. Trata-se de um experimento com estrutura de tratamento fatorial $2 \times 4 \times 4$ e estrutura de parcelas completamente aleatorizado. Os fatores estudados e seus respectivos níveis são: tratamento de desinfecção (TD) (aplicado ou não), tempo de secagem (0, 24, 72 e 360 horas) e temperatura de armazenamento (temperatura ambiente, 10, -20 e -196°C). Neste experimento, deseja-se avaliar o efeito dos tratamentos sobre a morte das sementes.

Para cada combinação dos níveis dos fatores, são alocadas 100 sementes e ao final do experimento foi contabilizado o número de sementes que morreram. Para cada um dos 32 tratamentos há 4 repetições, constituindo 128 observações no total.

Os dados foram modelados estatisticamente por meio de um modelo linear generalizado assumindo distribuição binomial para a proporção de sementes mortas. Para este conjunto de dados, são comparadas as estimativas do erro padrão dos parâmetros do modelo obtidos por máxima verossimilhança, quase-verossimilhança, *bootstrap* dos resíduos de Pearson e dos resíduos no preditor linear. O modelo final selecionado foi aquele contendo os efeitos principais dos fatores e as interações duplas Secagem:Armazenamento e Armazenamento:TD.

As estimativas dos parâmetros e os erros padrões obtidos para o modelo binomial, quase binomial, por *bootstrap* dos resíduos de Pearson padronizados e por *bootstrap* dos resíduos padronizados no preditor linear estão apresentados na Tabela 1.

Tabela 1: Estimativas dos parâmetros do modelo Binomial e seus erros padrões associados

Parâmetro	Estimativa	EP(Bin)	EP(Qbin)	EP_B(Pearson)	EP_B(PL)
(Intercepto)	-1.61	0.13	0.19	0.44	0.32
SEC24	-0.15	0.18	0.27	0.61	0.43
SEC72	-1.86	0.30	0.45	0.96	0.50
SEC360	-2.57	0.40	0.61	1.99	0.59
ARM-196	-2.04	0.30	0.44	1.12	0.51
ARM-20	-1.46	0.24	0.36	0.76	0.47
ARM10	-1.91	0.26	0.39	0.78	0.47
TDCOM	-1.83	0.22	0.33	0.56	0.35
SEC24:ARM-196	0.85	0.35	0.54	1.39	0.63
SEC72:ARM-196	1.61	0.48	0.74	1.90	0.72
SEC360:ARM-196	3.12	0.50	0.77	2.29	0.74
SEC24:ARM-20	0.10	0.31	0.47	0.96	0.61
SEC72:ARM-20	1.34	0.41	0.63	1.39	0.67
SEC360:ARM-20	2.40	0.48	0.73	2.23	0.74
SEC24:ARM10	0.56	0.31	0.47	0.97	0.60
SEC72:ARM10	1.69	0.41	0.63	1.48	0.68
SEC360:ARM10	2.30	0.49	0.76	2.23	0.75
ARM-196:TDCOM	1.43	0.31	0.47	0.80	0.48
ARM-20:TDCOM	1.68	0.29	0.44	0.78	0.46
ARM10:TDCOM	2.15	0.29	0.45	0.75	0.45

Verifica-se a presença de superdispersão (ou sobredispersão) nos dados, caracterizada pelo valor estimado do parâmetro de dispersão de $\hat{\phi} = 2.349$. Analisando-se os esquemas de reamostragem dos resíduos, observa-se que o *bootstrap* dos resíduos padronizados de Pearson gera erros padrões bastante elevados para algumas estimativas, como SEC360:ARM-196 e SEC360:ARM10. Pela Figura 1, tem-se a indicação de que isso ocorre devido à alguns valores estimados para estes parâmetros serem bastante extremos. Para estes mesmos parâmetros, pela Figura 2, observa-se que quando utilizados os resíduos padronizados no preditor linear, não são geradas estimativas tão discrepantes. Além disso, o segundo esquema, parece detectar uma variabilidade extra nas estimativas dos parâmetros de secagem e armazenamento, que não é comportada pelas estimativas de quase verossimilhança.

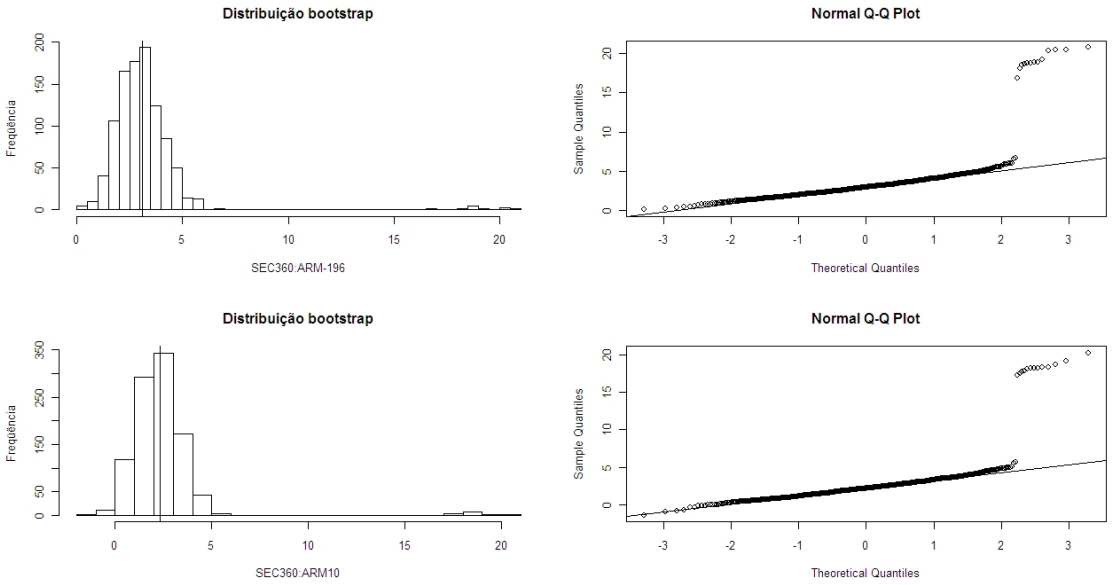

Figura 1: Histograma e Q-Q plot das estimativas *bootstrap* dos parâmetros SEC360:ARM-196 e SEC360:ARM10 - Resíduos de Pearson

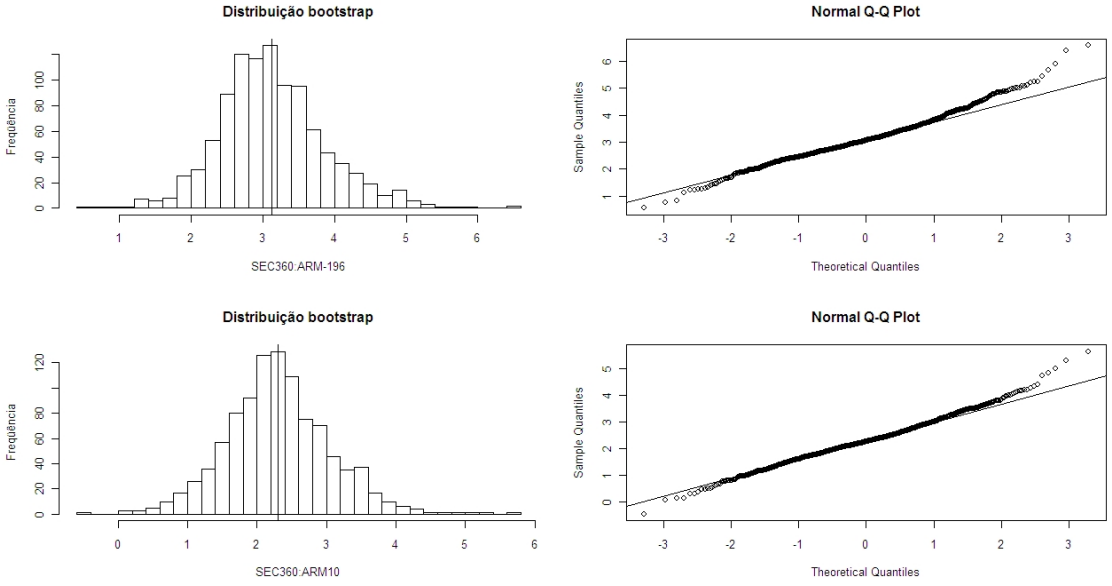

Figura 2: Histograma e Q-Q plot das estimativas *bootstrap* dos parâmetros SEC360:ARM-196 e SEC360:ARM10 - Resíduos no preditor linear

Referências

DAVISON, A.C.; HINKLEY, D.V. **Bootstrap methods and their application**. 1. ed. Nova Iorque, NY: Cambridge Press, 1997. 592p.

SMYTH, G. K. . **Generalized linear models with varying dispersion**. *Journal of the Royal Statistical Society B* **51** (1989), 47-60.