Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Planejamento ótimo de experimento para modelos de respostas binárias utilizando a teoria da decisão Bayesiana

Edmilson Rodrigues Pinto[1][2]
FAMAT/UFU, Uberlândia, MG

Cássio Antonio Andrade[3]
UFU, Uberlândia, MG

Ao planejar um experimento, as decisões devem ser tomadas antes da coleta de dados, a qual é restrita a recursos limitados. Como várias informações geralmente estão disponíveis antes do experimento ser realizado e, de fato, muitas vezes motivam a realização do experimento, os métodos Bayesianos são idealmente adequados para construção do planejamento experimental. A teoria da decisão Bayesiana também motiva a especificação precisa da razão pelo qual o experimento está sendo conduzido. A ideia básica em um projeto experimental é que a inferência estatística sobre as quantidades de interesse pode ser melhorada selecionando-se apropriadamente os valores das variáveis de controle [2].

O objetivo desta pesquisa é desenvolver um processo que permita encontrar o desenho experimental ótimo utilizando a metodologia Bayesiana, através da exploração da superfície de utilidade esperada via métodos de Monte Carlo para modelos lineares generalizados (MLG); em especial, modelos com respostas binárias.

O problema de obtenção de planejamento ótimo de experimentos dentro do contexto da teoria da decisão Bayesiana é formulado da seguinte forma. Seja um conjunto de m covariáveis, denominadas variáveis experimentais, dispostas em um vetor $\boldsymbol{x}^t = (x_1, ..., x_m)$. Os níveis dessas variáveis experimentais, para os quais as respostas $\boldsymbol{y}^t = (y_1, ..., y_n)$ serão observadas, são controlados pelo investigador e estão no interior de uma região pré-determinada, bem definida, um conjunto compacto $\mathcal{X}$, denominado região experimental. O valor esperado da variável resposta, associado a uma unidade experimental qualquer $\boldsymbol{x} \in \mathcal{X}$, $\eta(\boldsymbol{\beta}, \boldsymbol{x}) = \beta_1 f_1(\boldsymbol{x}) + \ldots + \beta_p f_p(\boldsymbol{x})$, é uma combinação linear de funções conhecidas de $\boldsymbol{x}$, em que $\boldsymbol{\beta}^t = (\beta_1, ..., \beta_p)$ é o vetor de parâmetros desconhecidos, tal que $\boldsymbol{\beta} \in \Omega$, onde Ω é o espaço paramétrico p-dimensional. O conjunto de vetores $\{\boldsymbol{x}_i, i = 1, \ldots, k\}$, representa uma possível escolha de k unidades experimentais, nas quais N respostas serão observadas. O desenho experimental, no caso discreto, é representado por $\xi = \{(x_1...x_k); (p_1...p_k)\}$, em que $p_j = \frac{r_j}{N}$, sendo r_j o número de replicações da unidade experimental j.

De acordo com [2], a obtenção do desenho ótimo Bayesiano pode ser colocada como um problema de decisão Bayesiana da seguinte forma. Especifica-se uma priori $\pi(\theta)$, definida sobre o espaço paramétrico Θ e uma função de utilidade $U(\theta, d^\xi, \xi, y)$ definida no espaço $\Theta \times \mathbb{D} \times \mathcal{C} \times \mathcal{Y}$, onde $\mathbb{D}$ é o espaço das decisões, $\mathcal{Y}$ é o espaço amostral e $\mathcal{C}$ é o conjunto de todas as medidas de probabilidade definidas sobre $\mathcal{X}$. A função $U(\theta, d^\xi, \xi, y)$ representa o ganho em observar $\boldsymbol{y}^t = (y_1, y_2, ..., y_n)$ com base no desenho ξ e tomando a decisão d^ξ. Nesse contexto, o objetivo é encontrar o desenho ξ^* que

[1]edmilson.pinto@ufu.br
[2]Agradecimento à FAPEMIG pelo apoio financeiro.
[3]andrade.cassio@hotmail.com

maximiza a utilidade esperada posterior $E_{\theta|y}[U(\theta, d^{\xi}, \xi, y)]$. Para maiores detalhes sobre a Teoria de Planejamento Ótimo de Experimentos veja [4].

O algoritmo proposto por [1], considerando o método de Metropolis-Hastings para atualizar $(\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{y})$, foi utilizado para obtenção do desenho ótimo. A utilidade considerada é baseada na informação de Shannon, a negativa da função de entropia. Desta forma, o critério para obtenção do desenho ótimo foi escolher o desenho que maximizasse o ganho esperado na informação de Shannon ou, equivalentemente, que maximizasse a distância esperada de Kulback-Leibler entre posteriori e priori. Veja [4] e [3] para mais detalhes.

O MLG considerado foi o binomial com ligação logística e preditor linear $\eta(x) = \beta_0 + \beta_1 x$. Para obtenção do desenho ótimo Bayesiano, foi especificado um desenho inicial ξ_0 com os pontos amostrais $(x_1 = -4; x_2 = 1; x_3 = 3)$ com seus respectivos pesos $(p_1 = 0,4; p_2 = 0,25; p_3 = 0,35)$. Para os parâmetros β_0 e β_1, foram consideradas as prioris normais $\beta_0 \sim \mathcal{N}(4; 2)$ e $\beta_1 \sim \mathcal{N}(0,2; 0,1)$. Considerando um tamanho amostral $n = 100$, foram gerados n valores para β_0 e n valores para β_1, conforme as distribuições a priori especificadas. Com os valores dos parâmetros simulados e do desenho inicial, calculamos os valores das respostas y_i, $i = 1 \ldots, n$. Com os valores simulados das respostas $\boldsymbol{y}$, calculamos o valor da entropia esperada via método de Monte Carlo. Geramos $N =$ 10000 simulações de Monte Carlo para cada um dos parâmetros β_0 e β_1. Para cada par (β_0, β_1), um vetor de resposta $\boldsymbol{y}$ de tamanho n foi simulado. Associado a cada vetor de respostas, considerando o desenho inicial escolhido e o par de parâmetros em questão, calculamos a verossimilhança, a qual juntamente com a priori, possibilitou a obtenção do cálculo da posteriori. Desta forma, obtivemos um vetor de posterioris de tamanho N. Pelo método de Monte Carlo, obtivemos a entropia esperada como a média do valor negativo do logaritmo do vetor de posterioris obtido.

O processo para gerar novos desenhos em busca daquele considerado ótimo, utilizou como distribuição proposta g uma distribuição normal centrada nos pontos de suporte do desenho experimental e com variância proporcional ao inverso dos pesos multiplicado pelo quadrado do produto dos pontos de suporte e de um coeficiente de variação (CV) igual a 0,2 , ou seja $g \sim \mathcal{N}(x, (CVx)^2 p^{-1})$. A razão entre a entropia esperada do desenho gerado e a entropia esperada do desenho corrente fornece uma probabilidade de aceitação no algoritmo. Esse processo foi repetido até atingir o critério de parada, que foi definido como sendo uma variação de 0,01% na média móvel dos 10 últimos valores das utilidades em relação aos 10 valores anteriores a esses últimos valores, porém apenas após 2000 iterações iniciais. Desta forma, o desenho ótimo obtido foi: $\xi = \begin{pmatrix} -0.05 & -0.03 & -0.02 & -0.01 & 0.00 & 0.01 & 0.02 & 0.03 & 0.04 & 0.06 \\ 0.18 & 0.07 & 0.10 & 0.06 & 0.06 & 0.03 & 0.07 & 0.07 & 0.15 & 0.21 \end{pmatrix}$.

Concluindo, verificamos que existe grande influência da priori e alta complexidade computacional na procura pelo desenho ótimo Bayesiano; e esta se torna ainda maior à medida que se aumenta a dimensão da região experimental. Assim, métodos eficientes de procura por desenhos experimentais ótimos em MLG é um tema bastante relevante e um desafio para pesquisas futuras.

Referências

[1] Bielza, C., Muller, P. e Rios-Insua, D. *Decision analysis by augmented probability simulation*, Management Science, v. 45, 995-1007, 1999.

[2] Chaloner, K., Verdinelli, I. Bayesian experimental design: a review. *Statistical Science*, v. 10, 273-304, 1995.

[3] Lindley, D. V. On a measure of information provided by an experiment. *The Annals of Mathematical Statististics* v. 27, 986-1005, 1956.

[4] Pinto, E. R., Ponce de Leon, A. C. M. *Planejamento Ótimo de Experimentos.* XVII Simpósio Nacional de Probabilidade e Estatística, Caxambu, Associação Brasileira de Estatística, 2006.

