COMPARAÇÃO DO AJUSTE DOS MODELOS DE REGRESSÃO CENSURADOS EM DIFERENTES PACOTES ESTATÍSTICOS

Rafael Cabral Fernandez[1], Gustavo Henrique Mitraud Assis Rocha[2] e Renata Souza Bueno[3]

Introdução

Os modelos de regressão censurados ou modelos Tobit são bastante usados na análise estatística para modelar variáveis respostas que são parcialmente observadas ou que possuem uma quantidade de valores agrupados em um valor limite. O campo de aplicação desses modelos cobre diversas áreas da ciência, tais como econometria, biometria, ensaios clínicos dentre outros. No trabalho pioneiro de Tobin (Tobin, 1958), o modelo Tobit assume que o valor limite é zero. Logo, apenas valores positivos da variável dependente são efetivamente observados. Todavia, a inferência no modelo Tobit leva em consideração todo o conjunto de dados, tanto as respostas efetivamente observadas, quanto as censuradas.

Para realizar a inferência dos parâmetros do modelo Tobit sob a abordagem bayesiana é necessária uma construção de algoritmos para o cálculo de integrais de alta dimensionalidade. Uma possibilidade para aproximar o valor de tais integrais é entendê-las como sendo valores esperados de uma determinada distribuição e considerar o método Monte Carlo via Cadeias de Markov (MCMC), cuja ideia é obter amostras e calcular estimativas amostrais da distribuição. Outra possibilidade é considerar o método de Laplace, que aproxima a distribuição alvo através de uma distribuição normal com vetor de médias igual ao vetor composto pela moda da densidade alvo e matriz de covariância igual à matriz hessiana da densidade da distribuição alvo. O desenvolvimento desse método, combinado com técnicas de integração numérica, resultam na abordagem INLA (Integrated Nested Laplace Approximations) (Rue et al., 2009). Em geral, essa abordagem gera resultados com um tempo de execução menor quando comparado com os métodos MCMC (Rue et al., 2017; Meehan et al., 2018). Em termos metodológicos, o MCMC é um procedimento estocástico enquanto o INLA é uma aproximação analítica.

[1] Escola Nacional de Ciências Estatísticas (ENCE), rafael.fernandez@fgv.br

[2] Escola Nacional de Ciências Estatísticas (ENCE), gustavo.rocha@ibge.com.br

[3] Escola Nacional de Ciências Estatísticas (ENCE), renata.bueno@ibge.com.br

Objetivos

A comparação entre as metodologias será pautada em critérios como estatísticas resumo, precisão, tempo computacional, entre outros. Serão considerados modelos lineares sem censura e censurados. Em linhas gerais, é de interesse verificar o desempenho de métodos de aproximação analítica, em contraposição aos métodos de simulação estocástica, já bem implementados na literatura.

Material e Método

Os modelos utilizados podem ser descritos da seguinte forma:

$$\eta_i = \beta_0 + \beta_1 x_i + \epsilon_i \;\; ; \;\; \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

$$y = \begin{cases} \eta, \; se \;\; \eta \geq \tau_E \\ \xi, \; se \;\; \eta < \tau_E, \end{cases} \tag{2}$$

A equação (1) representa um modelo de regressão linear simples, considerando $\boldsymbol{\beta_0}$ como intercepto, $\boldsymbol{\beta_1}$ como coeficiente angular e $\boldsymbol{\epsilon_i}$ como os erros, supostos normalmente distribuídos, com média zero e variância constante. Ainda em (1), $\boldsymbol{x_i}$ denota a variável explicativa, geradas de uma distribuição normal com média 0 e variância 1, já $\boldsymbol{\eta_i}$ denota a variável dependente a ser explicada. A equação (2) pressupõe um modelo com dados censurados à esquerda, sendo $\boldsymbol{\eta}$ o valor real, censurado por um limitante $\boldsymbol{\tau_E}$, de tal forma que o modelo registra um valor de referência $\boldsymbol{\xi}$, indicando que houve o fenômeno de censura.

Foram definidas as seguintes distribuições a priori para os parâmetros do modelo linear definido em 1: $\beta_0 \sim N(0{,}100)$; $\beta_1 \sim N(0{,}100)$; $\tau \sim Gamma(1; 0{,}1)$, onde $\tau = (\sigma^2)^{-1}$, isto é, o parâmetro de precisão do erro. De modo que E(τ) = 10 e Var(τ) = 100.

Para a execução dos métodos MCMC serão consideradas as ferramentas WinBUGS (Spiegelhalter et al., 1995), JAGS (Plummer, 2016), Stan (Stan Development Team, 2015) e NIMBLE (de Valpine et al., 2017) através do software R (R Core Team, 2018). A abordagem INLA será considerada através do pacote R-INLA (Martins et al., 2013) do R. A comparação final se dará, portanto, para os 5 softwares apresentados, mais uma versão da aproximação de Laplace, implementada manualmente.

Para efeito de comparabilidade entre os modelos gerados para cada pacote proposto, foi fixado uma semente aleatória. A partir de um modelo inicial, definido $Y_i = -2 + 2x_i + \epsilon_i$, onde $\epsilon_i \sim N(0{,}5)$. Foram gerados 100 conjuntos de 100 observações cada. Para cada conjunto, foi ajustado um modelo de regressão linear simples, apresentado em (1). Cada uma das 5 propostas computacionais (Bugs, Jags, Stan, Nimble e Laplace) foi comparada mediante as seguintes estatísticas: Média a posteriori, desvio padrão a posteriori, erro

quadrático médio (utilizando a média a posteriori como estimador), intervalo de cobertura e tempo computacional gasto.

Resultados e Discussão

A Tabela 1 apresenta os resultados obtidos a partir das comparações propostas. O objetivo inicial de cada pacote apresentado é estimar a maior quantidade de parâmetros com a maior precisão (menor erro quadrático médio) e o menor custo computacional possível. Verifica-se que a aproximação de Laplace apresenta o menor custo computacional, por outro lado, o Nimble apresenta o maior custo computacional. A precisão para qualquer uma das propostas é similar. O intervalo de cobertura para o INLA é considera ineficiente.

Tabela 1 – Resultados preliminares para 100 simulações de um modelo de regressão linear de uma variável explicativa

Pacote	Parâmetro	Média	Desvio Padrão	EQM	Cobertura (%)	Tempo (segundos)
Bugs	β_0	-2,0118	0,2252	0,0690	87%	250
	β_1	2,0204	0,2249	0,0470	96%	
	τ	0,2078	0,0283	0,0009	94%	
Jags	β_0	-1,9242	0,2198	0,0689	90%	48
	β_1	2,0261	0,2256	0,0482	94%	
	τ	0,2075	0,0283	0,0009	96%	
Stan	β_0	-2,0209	0,2245	0,0704	89%	84
	β_1	2,0318	0,2256	0,0486	97%	
	τ	0,2078	0,0283	0,0009	96%	
Nimble	β_0	-2,0122	0,2304	0,0692	89%	4342
	β_1	2,0238	0,2252	0,0468	94%	
	τ	0,2079	0,0283	0,0009	95%	
Laplace	β_0	-1,9941	0,2249	0,0703	96%	1
	β_1	1,9768	0,2249	0,0479	96%	
	τ	0,2071	0,0300	0,0011	98%	

Fonte: dados simulados

Conclusão

Para modelos que não apresentam estruturas complexas, como a regressão linear, tanto métodos estocásticos quanto aproximações analíticas fornecem resultados precisos, embora apresentem diferenças no tempo computacional de execução. Em linhas gerais,

para modelos lineares, metodologias com base estocástica tendem a apresentar resultados ligeiramente mais precisos (com menor variância) enquanto metodologias baseadas em aproximações analíticas tem um considerável ganho em tempo computacional. Existe então a necessidade de se adotar alguma medida de *trade-off*. O trabalho aqui apresentado, no presente momento, ainda se encontra em desenvolvimento e está aplicando o procedimento proposto para modelos de dados censurados.

Referências

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., Rastislav Bodik, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. Journal of Computational and Graphical Statistics, 26, (2), 403-413.

Martins, T. G., Simpson, D., Lindgren, F., Rue, H. (2013). Bayesian computing with INLA: New features. Computational Statistics & Data Analysis, 67, 68-83.

Meehan, T. D., Michel, N. L., Havard, R. (2018). Estimating animal abundance with N-mixture models using the R-INLA package for R. ARXIV. arXiv:1705.01581.

Plummer, M. (2016). rjags: Bayesian Graphical Models using MCMC. R package version 4-6.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B. 71, (2), 319-392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., Lindgren, F. K. (2017). Bayesian computing with NLA: a review. Annual Review of Statistics and its Application. 4, 395-421.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R. (1995). BUGS: Bayesian inference using Gibbs sampling. Version 0.50, MRC Biostatistics Unit, Cambridge.

Stan Development Team (2015). Stan Modeling Language User's Guide and Reference Manual. Version 2.6.1.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.