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).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Análise do método MI-LASSO com validação cruzada para dados de COVID-19

Diego S. Santos[1]
CCEN/UFPE, Recife, PE
Pablo M. Rodriguez[2]
CCEN/UFPE, Recife, PE
Luz M. Gómez G.[3]
HC-FM-USP, São Paulo, SP
Heraldo P. de Souza[4]
HC-FM-USP, São Paulo, SP

1 Introdução

Em inúmeros estudos envolvendo análise de regressão é comum a presença de dados faltantes, principalmente em ensaios clínicos e na área médica no geral. Uma técnica que vem se tornando comum nos últimos anos é a imputação múltipla (IM) que preenche os dados faltantes a partir das informações do próprio banco de dados por meio de alguma regra específica gerando vários bancos de dados imputados. Quando tem-se muitas variáveis de interesse para se trabalhar, uma técnica bastante popular é o *Least Absolute Shrinkage and Selection Operator* (LASSO), sendo este um método de regularização; i.e., técnica que desencoraja o ajuste excessivo dos dados, afim de diminuir a sua variância. Um problema que surge quando se trabalha com o LASSO em dados multi-imputados é quanto a função de penalização, pois para cada conjunto de dados imputado é possível que se tenha a escolha de diferentes covariáveis para cada conjunto de dados o que pode prejudicar a escolha de quais covariáveis manter no modelo. Uma técnica que resolve este problema é o MI-LASSO, este contendo uma função de penalização conjunta para forçar a escolha das mesmas covariáveis nos conjuntos de dados imputados. Neste trabalho realizamos uma análise do método MI-LASSO com validação cruzada para dados de COVID-19, nome dado à doença causada pelo vírus SARS-CoV-2.

2 Metodologia

Seja X_n a matriz $n \times p$ de covariáveis, e Y_d o vetor $n \times 1$ de variáveis resposta para o d-ésimo conjunto imputado, $d \in \{1, \ldots, D\}$, em que D é a quantidade de conjuntos imputados. Sejam $X_{d,i}$ o vetor da covariável $p \times 1$ para a i-ésima observação no d-ésimo conjunto imputado, $Y_{d,i}$ a resposta para a i-ésima observação no d-ésimo conjunto imputado, e $X_{d,j}$ a j-ésima covariável para a i-ésima observação no d-ésimo conjunto imputado. O vetor $p \times 1$ de coeficientes para o d-ésimo conjunto de

[1]diego.silva@ufpe.br
[2]pablo@de.ufpe.br
[3]luzgomez@alumni.usp.br
[4]heraldo.possolo@fm.usp.br

dados é dado por β_d, o coeficiente no d-ésimo conjunto correspondente à j-ésima covariável é dado por $\beta_{d,j}$, e o intercepto para o d-ésimo conjunto de dados é dado por μ_d. O vetor de parâmetros de regressão para o d-ésimo conjunto de dados é dado por $\theta_d = (\mu_d, \beta_d)$. A função conhecida como MI-LASSO proposta originalmente em [1] é dada por

$$(\hat{\theta}_1, \ldots, \hat{\theta}_D) = \underset{\theta_1,\ldots,\theta_D}{\arg\min} \left\{ -\frac{1}{n} \sum_{d=1}^{D} \sum_{i=1}^{n} \log L(\theta_d | Y_{d,i}, X_{d,i}) + \lambda \sum_{j=1}^{p} \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2} \right\}, \tag{1}$$

em que $\lambda \in [0, \infty)$ é um parâmetro de ajuste, e a função de penalização conjunta é conhecida como Lasso de grupo (GLASSO). Embora os θ_d's não sejam idênticos, para qualquer j fixo, a penalidade de grupo Lasso reduz conjuntamente todos os $\beta_{d,j}$'s a zero; i.e, $\beta_{1,j} = \cdots = \beta_{D,j} = 0$. Com base em [1], consideramos a seguinte penalidade

$$P(\beta_1, \ldots, \beta_D) = \sum_{j=1}^{p} \hat{a}_j \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2}, \tag{2}$$

em que $\hat{a}_j = \left(\sqrt{\sum_{d=1}^{D} \hat{\beta}_{d,j}^2} + 1/(nD) \right)^{\gamma}$, $\hat{\beta}_{d,j}$ é estimado pelo GLASSO, $\gamma = \lceil 2v/1 - v \rceil + 1$, e $v = \log(pD)/\log(nD)$. A equação (1) com penalidade dada em (2) é conhecida como Lasso de grupo adaptativo (GALASSO) e sua implementação no software `R` é dada pela função `galasso` do pacote `myselect`.

3 Objetivos e primeiros resultados

Neste trabalho avaliamos o impacto do uso da técnica de validação cruzada na seleção de variáveis no MI-LASSO em relação a implementação usual e consequentemente seu impacto na predição, dada pela função `cv.galasso` do mesmo pacote `myselect`. Aplicamos as técnicas a dados de pacientes da Covid-19, fornecidos pelo Grupo de Emergências do Hospital das Clinicas da USP. Especial ênfase é dada ao índice de oxigenação no sangue como variável resposta. Nossos resultados preliminares mostram que há um impacto considerável na seleção das covariáveis quando se é utilizada a validação cruzada, esta exclui mais covariáveis em relação ao modelo sem a validação. Em particular, as seguintes variáveis são estatisticamente significativas dentro do modelo: pressão sistólica na admissão, pressão diastólica na admissão, peso na admissão (kg), hemoglobina 72h e tempo de intubação total. Isto indica que estas variáveis ajudam a explicar a variabilidade do índice de oxigenação em sangue do paciente.

4 Agradecimentos

Os autores agradecem à FAPESP, processo número 16/14566-4, pelo auxílio financeiro.

Referências

[1] Chen, Qixuan, and Sijian Wang. *Variable selection for multiply-imputed data with application to dioxin exposure study.* Statistics in Medicine 2013 Sep 20;32(21):3646-59. DOI: 10.1002/sim.5783.

[2] Du, Jiacong, et al. *Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods.* arXiv preprint arXiv:2003.07398 (2020).