Verão IDOR – ImpaTech
Tema: Modelagem e inferência estatística em alta dimensão
Professor: Rodney Fonseca
Datas: 12 e 14 de janeiro de 2026

Exercícios propostos

1. Este exercício utilizará os dados de mamografia (Fan et al., 2020, p. 254) vistos em aula.

   (a) Refaça a análise dos dados de mamografia usando preditores padronizados:

   $$x_{ij}^p = \frac{x_{ij}}{\sqrt{n^{-1}\sum_{i=1}^n x_{ij}^2}} = \frac{x_{ij}}{s_j}.$$

   Durante o ajuste do modelo, utilize a opção `standardize='FALSE'` no `glmnet`.

   (b) Interprete estimativas para a idade levando o fator de escala s_j em consideração.

2. Considere agora os dados sobre aleitamento materno analisados em Colosimo and Giolo (2006, p. 140). Os dados também estão disponíveis no arquivo `desmame.txt` em https://docs.ufpr.br/~giolo/Livro/ApendiceA/Arquivos_ASA.html. O conjunto de dados tem as seguintes características:

   - a variável resposta é o tempo máximo de aleitamento materno
   - a variável indicadora vale 1 se o tempo do desmame foi observado e 0 caso contrário
   - as onze variáveis explicativas são categóricas e tem nomes V1, V2, ..., V11

   O objetivo é escolher um subconjunto de variáveis para modelar o tempo de aleitamento materno. Faça uma análise descritiva dos dados e ajuste o modelo de Cox penalizado com lasso incluindo todas interações entre covariáveis ($p = 77$).

3. Neste exercício serão analisados os dados de cardiomiopatia disponíveis na página das aulas. A variável resposta é uma medida de uma proteína relacionada à doença de cardiomiopatia (primeira coluna) e as covariáveis são $p = 6319$ expressões genéticas (demais colunas). As medidas foram coletas de $n = 30$ ratos. O objetivo é identificar quais destes genes são mais relevantes para modelar a variável resposta. Mais detalhes sobre o banco de dados podem ser conferidos em (Fan et al., 2020, p. 429) e na página destes autores https://runzelipsu.github.io/DataScience/ (*Cardiomyopathy Microarray Data*).

   (a) Obtenha estimativas usando o lasso e o lasso deviesado.

   (b) Construa intervalos de 95% de confiança usando o lasso deviesado e verifique quais destes não contém zero.

   (c) Verifique quais covariáveis possuem valor-p abaixo de $\alpha = 0.05/p = 7.9113 \cdot 10^{-6}$ (correção de Bonferroni).

   (d) Construa intervalos de 95% de confiança bootstrap e compare os resultados com as conclusões obtidas no item (b). O conjunto de variáveis cujo intervalo não contém o zero foi o mesmo?

Referências

E. A. Colosimo and S. R. Giolo. *Análise de sobrevivência aplicada*. Blucher, São Paulo, 2006.

J. Fan, R. Li, C.-H. Zhang, and H. Zou. *Statistical Foundations of Data Science*. CRC Press, Boca Raton, 2020.