UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Diego da Silva Santos

Modelos de regularização com imputação e curvas de decisão aplicados a dados de medicina

Recife

2022

Diego da Silva Santos

Modelos de regularização com imputação e curvas de decisão aplicados a dados de medicina

Trabalho apresentado ao programa de pós-graduação em Estatística do departamento de Estatística da universidade federal de pernambuco como requisito parcial para obtenção do grau de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Pablo Martín Rodríguez

Coorientador (a): Luz Marina Gómez Gómez

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

S237m Santos, Diego da Silva
Modelos de regularização com imputação e curvas de decisão aplicados a dados de medicina / Diego da Silva Santos. – 2022.
64 f.: il., fig., tab.

Orientador: Pablo Martín Rodríguez.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.
Inclui referências.

1. Estatística aplicada. 2. Imputação múltipla. 3. Regressão regularizada. 4. Validação cruzada aninhada. 5. Curvas de decisão. I. Martín Rodríguez, Pablo (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2022 - 82

DIEGO DA SILVA SANTOS

MODELOS DE REGULARIZAÇÃO COM IMPUTAÇÃO E CURVAS DE DECISÃO APLICADOS A DADOS DE MEDICINA

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 22 de fevereiro de 2022.

BANCA EXAMINADORA

Prof. Dr. Pablo Martín Rodriguez
DE/UFPE

Profª. Drª Florencia Graciela Leonardi
USP

Profª. Drª Tatiana Andrea Benaglia Carvalho
UNICAMP

Dedico está dissertação a minha família que sempre tem me apoiado em toda minha trajetória.

AGRADECIMENTOS

Agradeço primeiramente a DEUS por me permitir trilhar o caminho até aqui sempre amparado pela sua infinita misericórdia. Minha família nunca deixou de me apoiar um segundo se quer e sempre fizeram tudo para que eu pudesse ter oportunidades na vida, e esta é uma delas, eu os agradeço por todo o sempre e dedico todo este trabalho a minha mãe Elza e irmãos Herbson, Felipe e Anderson, e ao meu querido pai que já se foi, de onde estiver eu o saúdo.

Aos amigos que estão sempre comigo me apoiando e se orgulhando de minhas conquistas e crescimento, em especial aos meus melhores amigos Bruno Fernandes e Ingrid Gomes, e outra gama de amigos que não são do meu âmbito acadêmico que ficaria demasiado citar a todos. Partindo para os amigos que fiz na graduação que hoje são parceiros de vida cito aqui Áurea Fonseca, João Victor, Lucas Araújo, Daniele Norões, Alberto Rodrigues, Yohana Gonçalves, Matheus Osterno e os demais que não citei mas que sempre contribuem com a amizade e companheirismo.

Partindo para o âmbito acadêmico, agradeço ao meu professor orientador Pablo Martín Rodríguez e professora coorientadora Luz Marina Goméz Goméz pelo apoio, paciência e aprendizado. Cito ainda o professor Luiz Gustavo de Bastos Pinho que foi meu orientador na graduação e hoje é um amigo sempre disposto a ajudar.

Por fim, agradeço a agência de fomento CAPES pelo apoio financeiro e ao PPGE-UFPE pela oportunidade de fazer parte deste conjunto de pesquisadores, alunos e profissionais.

RESUMO

Na análise estatística é comum a presença de dados faltantes em muitas aplicações e estudos em inúmeras áreas com especial enfase a área da saúde. Estudos foram sendo desenvolvidos ao longo da segunda metade do século XX para contornar o problema de dados faltantes dos quais destacam-se os trabalhos de RUBIN (1988) e SCHAFER (1997) para imputação de darubin1988overviewos. Além do tratamento do banco de dados e preenchimento dos dados faltantes para utilização das técnicas estatísticas de modelagem, que em sua grande maioria são restritas a dados completos, outra questão que se levanta após a imputação é a técnica estatística mais adequada a ser utilizada para o determinado objetivo inferencial. Na área de regressão os modelos com regularização vem sendo cada vez mais utilizados em problemas de alta dimensão onde tem-se muitas covariáveis a serem estimadas ou em problemas envolvendo multicolinearidade. Esta dissertação aborda o problema da modelagem de regressão regularizada aplicada aos dados imputados, em especial ao modelo de regressão LASSO adaptativo para dados multi-imputados conhecido como GALASSO (DU et al., 2020), também utiliza-se a técnica de validação cruzada aninhada (BATES; HASTIE; TIBSHIRANI, 2021) para obtenção da variância empírica de validação cruzada e intervalos de confiança mais largos para o erro de validação dentro da amostra envolvidos nos modelos de regularização. Desta forma, é proposta uma abordagem utilizando a imputação múltipla através do Bootstrap Bayesiano atrelado ao modelo LASSO logístico com validação cruzada aninhada para correção da estimativa de variância e intervalo de confiança da validação cruzada usual, buscando-se o melhor poder de classificação. Por fim, utiliza-se da metodologia de curvas de decisão proposta por VICKERS; ELKIN (2006) para a aplicação em dados de COVID-19 com o intuito de propor uma abordagem correta na tomada de decisões de profissionais da saúde em problemas de diagnóstico clínico na presença de dados faltantes.

Palavras-chaves: imputação múltipla; regressão regularizada; validação cruzada aninhada; curvas de decisão.

ABSTRACT

In statistical analysis, the presence of missing data is common in many applications and studies in numerous areas, with special emphasis on health. Studies were developed throughout the second half of the XX century to overcome the problem of missing data, of which the works of RUBIN (1988) and SCHAFER (1997) for data imputation stand out. In addition to processing the database and filling in the data for the use of statistical modeling techniques, which are mostly restricted to complete data, another issue that arises after processing the data is the most appropriate statistical technique to be used for the given inferential objective. In the area of regression, models with regularization have been increasingly used in high-dimensional problems where there are many covariates to be estimated or multicollinearity problems. This dissertation addresses the problem of regularized regression modeling applied to imputed data, especially to the adaptative LASSO regression model for multi-atput data known as GALASSO (DU et al., 2020), also using the nested cross-validation technique (BATES; HASTIE; TIBSHIRANI, 2021) to obtain the empirical variance of cross-validation and wider confidence intervals for the in-sample validation error involved in the regularization models. Thus, an approach is proposed using multiple imputation through Bayesian Bootstrap linked to the logistic LASSO model with nested cross-validation to correct the variance estimate and confidence interval of the usual cross-validation, seeking the best classification power. Finally, the methodology of decision curves proposed by VICKERS; ELKIN (2006) is applied to COVID-19 data in order to propose a correct approach to decision-making by health professionals in clinical diagnosis problems in the presence of missing data.

Keywords: multiple imputation; regularized regression; nested cross validation; decision curves.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A área de aprendizagem de maquina (*machine learning*) vem sendo aprimorada nos últimos anos e tem sido aplicada em inúmeras áreas do conhecimento, principalmente quando se trabalha com grandes quantidades de dados, conhecidos como *Big Data*. Quando tem-se interesse em modelar fenômenos a partir de determinadas variáveis conhecidas, que de alguma maneira regem o fenômeno em questão, recaímos em um problema de regressão, ou seja, explicar determinado resultado que denotamos por variável resposta em decorrência de uma ou mais variáveis que chamamos de explicativas. Adicionalmente, quando o número de variáveis explicativas é muito alto, maior do que o número de observações, ou quando estas variáveis são correlacionadas, temos um problema envolvendo alta dimensão ou multicolinearidade, que podem ser contornados com o uso de modelos de regressão regularizados. Na aprendizagem de máquina, utiliza-se do conceito de regressão porém com um foco em predizer resultados sem se focar muito na questão conceitual e dinâmica do fenômeno, ou seja, na prática estatística padrão o interesse maior estar na parte conceitual e inferencial da modelagem enquanto que em aprendizagem de máquina geralmente foca-se mais no ganho preditivo e o quão bom esta predição se mostra (IZBICKI; SANTOS, 2020).

Inicialmente as técnicas estatísticas como análise de regressão foram pensadas e construídas para lidar com determinado padrão de dados, este padrão sendo matrizes completas $n \times p$, n sendo a quantidade de observações disponíveis e p a quantidade de variáveis envolvida no fenômeno estudado (RUBIN, 1988). Na prática, em muitos casos e estudos não temos total controle sobre os dados, de forma que muitas vezes não dispomos de tabelas com um padrão totalmente preenchido. Frente a isso, pouco mais da segunda metade do século passado, foi-se pensando em formas de lidar e se trabalhar com dados incompletos. Inicialmente excluía-se os dados e desta forma utilizava-se das técnicas estatísticas com os dados completos. Posteriormente viu-se que não era o ideal visto que, dependendo da quantidade de dados faltantes, perdia-se muita informação podendo assim chegar-se a resultados não condizentes com a real dinâmica do fenômeno em questão. Então pensou-se em um preenchimento destes dados por meio de alguma técnica, levando em consideração a informação que se tinha em mãos, o que foi chamado de imputação de dados. Foram surgindo então várias técnicas de preenchimento como imputação pela média, regressão, regressão estocástica entre outras.

O problema na abordagem deste tipo de imputação, que posteriormente foi chamada de

imputação única, era justamente o fato de inserir dados fictícios de forma única e isso podia enviesar de forma significativa os resultados e estes podiam não condizer e ficarem muito destoantes da realidade. Uma forma de diminuir este viés envolvido em preencher os dados uma única vez é a imputação múltipla proposta por RUBIN (1981), que ao longo das décadas de 80 e 90 foram sendo publicados inúmeros trabalhos para seu aprimoramento (RUBIN, 1981; RUBIN; SCHENKER, 1986; RUBIN, 1987; RUBIN, 1988; SCHAFER, 1997). O método consiste no preenchimento de dados faltantes com alguma técnica de imputação de forma iterativa com o intuito de construir vários bancos imputados, quantos forem necessários, para se ter vários bancos hipotéticos que de alguma forma poderiam ter sido observados buscando a diminuição de viés associada a imputação única.

Uma das áreas onde mais se tem que trabalhar com dados faltantes é a área da saúde, tanto em estudos epidemiológicos, diagnóstico e tratamento de doenças, etc. A aprendizagem de máquina nos últimos anos vem sendo largamente utilizada na área médica, principalmente em prognósticos de doenças onde o médico toma decisões sobre seus pacientes, e desta forma um modelo preditivo pode ser usado para auxiliar o médico na tomada de decisão de diagnósticos e tratamentos clínicos. Para um problema de tomada de decisão de diagnóstico tem-se um problema de classificação onde o médico precisa diagnosticar presença ou não de doença e desta forma iniciar ou não um tratamento. Existem inúmeros modelos de classificação que podem ser utilizados pelos profissionais e existem formas de decidir entre estes modelos de classificação. Uma das formas mais conhecidas de decidir entre modelos de clasśificação é utilizando as curvas ROC (*Receiver Operating Characteristic*) e o AUC (*Area Under The Curve*), sendo uma forma visual e uma métrica para decidir entre os modelos (SWETS, 1988).

A questão problemática do uso da curva ROC e medida de AUC é que elas apenas decidem que modelo desempenhou um melhor papel na classificação do diagnóstico, porém, em algumas situações as decisões médicas sobre diagnosticar um paciente envolve consequências para o paciente como tratamentos de câncer por exemplo. Uma abordagem proposta por VICKERS; ELKIN (2006) para contornar este problema foram as curvas de decisão, estas incorporam as consequências das decisões tomadas de prognósticos.

Temos então um problema que envolve uma forma adequada de utilização de base de dados incompleta, modelagem e classificação a partir destes dados e escolha do melhor modelo e prognóstico. Neste trabalho utilizamos de técnicas adequadas envolvendo todas estas abordagens em conjunto para chegar a resultados coerentes e condizentes de forma a ajudar da melhor maneira possível na tomada de decisões de profissionais da saúde para este

tipo de problema.

Nos capítulos 2 a 4 realizamos uma revisão bibliográfica e apresentamos os conceitos ligados a dados faltantes e imputação de dados, modelos de regressão em alta dimensão ou na presença de multicolinearidade, os chamados modelos de regressão regularizados como *LASSO*, *Rigde* e *Elastic Net*, estes modelos sendo aplicados a imputação de dados dos quais os trabalhados nesta dissertação é o GALASSO (*Multiple imputation grouped adaptive LASSO*) e o SAENET (*Stacking adaptive elastic net*), a metodologia de validação cruzada aninhada e os métodos de medida de desempenho e escolha de modelos de classificação para diagnósticos como curvas ROC, AUC e curvas de decisão.

No Capítulo 5 realizamos simulações e uma aplicação em dados reais de Covid-19 fornecido pelo Hospital das Clínicas da USP, estes estando em dois cenários diferentes, quando $p > n$ e na presença de multicolinariedade. A motivação para nossas avaliações estão em responder algumas questões: (i) como a retirada de dados pode impactar nas estimativas de viés do erro de validação cruzada aninhada em relação ao erro de validação cruzada, a estimativa de desvio padrão do erro de validação e a construção adequada de intervalos de confiança para o erro; (ii) como cada abordagem se comporta referente ao poder preditivo em relação a várias medidas de desempenho e na decisão final do profissional de saúde pela curva de decisão. Todos os resultados são apresentados e comentados com breves discussões.

2 DADOS FALTANTES

Em inúmeras áreas do conhecimento é comum se trabalhar com bases de dados que apresentam dados faltantes, principalmente na área da saúde no geral por apresentar coleta de dados por exemplo por meio de preenchimento de formulários dos pacientes e dados clínicos, comumente realizado em estudos epidemiológicos (MIOT, 2019; NUNES; KLÜCK; FACHEL, 2009). Esta dificuldade é bem conhecida e é necessário cuidado ao utilizar estas bases de dados para fins estatísticos. Naturalmente a primeira forma de contornar o problema com dados deste tipo foi a análise de caso completo que consiste na exclusão dos valores faltantes, utilizando assim o banco reduzido para as análises e inferências. Este tipo de abordagem pode gerar inconsistências dos resultados obtidos pois estes podem sofrer com viés devido a exclusão de informação, já que não há garantias do grau de impacto que os dados faltantes teriam sobre as conclusões do estudo em questão (LITTLE; RUBIN, 2019). Uma técnica que busca contornar este problema é a imputação de dados. De maneira formal, dados faltantes são valores não observados que seriam significativos para análise caso fossem observados, ou seja, um valor ausente que oculta um valor significativo. É necessário que faça sentido pensar que existem valores subjacentes reais que teriam sido observados se o levantamento dos dados tivesse sido melhor, ou não tivessem ocorridos problemas inerentes ao pesquisador quanto a coleta. Desta forma, quando a característica subjacente faz sentido, trabalha-se com analises de imputação de dados em que os dados faltantes são preenchidos por meio de alguma regra para se obter um banco de dados com matrizes de entrada totalmente preenchidas com a qual se possa trabalhar com técnicas estatísticas usuais.

2.1 MECANISMOS DE NÃO RESPOSTA

Uma etapa de extrema importância quando se trabalha com dados faltantes é a definição do mecanismo de não resposta envolvido no fenômeno em estudo. O primeiro mecanismo que definimos é o MAR (*missing at random*), ausente (faltante) ao acaso), que caracteriza uma perda onde pode-se rastrear os dados faltantes aos seus valores subjacentes que poderiam ter sido observados. Além disso, uma outra suposição é necessária para MAR: os parâmetros que regem o mecanismo de dados perdidos e os parâmetros que regem o parâmetro do fenômeno analisado devem ser "distintos" (a distribuição que rege a geração de dados faltantes é dife-

rente da distribuição que rege a variável de interesse). Sendo satisfeitas as duas condições, o mecanismo de perda dos dados pode ser ignorado (LITTLE; RUBIN, 2019).

Seja $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{falt})$ uma matriz $n \times p$ composta dos dados observados $\mathbf{Y}_{obs}$ e dos dados faltantes $\mathbf{Y}_{falt}$, e seja $\mathbb{I} = [\tau_{i,j}]$ com $i = (1, \ldots, n)$ e $j = (1, \ldots, p)$ uma variável indicadora tal que $\tau_{i,j} = 1$ para $y_{i,j}$ faltante e $\tau_{i,j} = 0$ para $y_{i,j}$ observado. Considerando $\tau_{i,j}$ como uma variável aleatória podemos definir uma distribuição subjacente que rege o mecanismo de falta dos dados. Seja $f(\mathbb{I}|\mathbf{Y}, \phi)$ a distribuição condicional de $\mathbb{I}$ dado $\mathbf{Y}$ com ϕ indicando parâmetros desconhecidos que regem o mecanismo de perda dos dados. Se o mecanismo de ausência de dados não depender dos dados $\mathbf{Y}$ a distribuição se reduz a

$$f(\mathbb{I}|\mathbf{Y}, \phi) = f(\mathbb{I}, \phi), \ \forall \ \mathbf{Y}, \ \phi.$$

Este caso é conhecido como MCAR (*missing completely at random*, faltante completamente ao acaso). Um exemplo simples para ilustração pode ser de uma fábrica onde uma maquina meça o comprimento de peças de uma linha de produção, e, por algum problema, esta máquina em certas peças não mediu o comprimento por falha mecânica. Este tipo de falta de dados pode ser considerada puramente estocástica se a máquina a priori não apresentou defeitos e estava em perfeito estado.

Considerando que a distribuição condicional de $\mathbb{I}$ depende apenas da parte observada dos dados $\mathbf{Y}_{obs}$ temos

$$f(\mathbb{I}|\mathbf{Y}, \phi) = f(\mathbb{I}|\mathbf{Y}_{obs}, \phi), \ \forall \ \mathbf{Y}_{falt}, \ \phi.$$

Esta outra forma é a conhecida MAR que desempenha um papel importante nos trabalhos com imputação de dados.

Voltemos a suposição de distinção dos parâmetros da geração dos dados faltantes ϕ e os parâmetros θ que regem a distribuição de Y. De uma perspectiva bayesiana, considerando independência dos parâmetros que regem o mecanismo de perda e os parâmetros da distribuição da variável resposta, podemos separar a distribuição priori conjunta no produto das prioris marginais

$$\Psi(\theta, \phi) = \Psi(\theta)\Psi(\phi). \tag{2.1}$$

Desta forma, não precisamos modelar o mecanismo de dados perdidos (mecanismo de perda ignorado) para estimar θ. Satisfeita as suposições de MAR e distinção podemos escrever

a verossimilhança dos dados observados da seguinte forma

$$\begin{aligned} f(\mathbf{Y}_{obs}, \mathbb{I}|\theta, \phi) &= \int f(\mathbb{I}|\mathbf{Y}, \phi) f(\mathbf{Y}|\theta) d\mathbf{Y}_{falt} \\ &= f(\mathbb{I}|\mathbf{Y}_{obs}, \phi) \int f(\mathbf{Y}|\theta) d\mathbf{Y}_{falt} \\ &= f(\mathbb{I}|\mathbf{Y}_{obs}, \phi) f(\mathbf{Y}_{obs}|\theta). \end{aligned} \tag{2.2}$$

Segundo SCHAFER (1997), esta fatoração só é possível com a suposição de MAR visto que θ pertencia originalmente aos parâmetros do modelo de dados completos. A partir da fatoração podemos descartar o termo que não depende de θ que é o parâmetro alvo de inferência, isto sendo possível pela suposição de (2.1). Desta forma

$$f(\mathbf{Y}_{obs}|\theta) \equiv L(\theta; \mathbf{Y}_{obs}), \tag{2.3}$$

é a verossimilhança dos dados observados. A equação (2.2) é conhecida como verossimilhança total dos dados observados e (2.3) é conhecida como verossimilhança simples dos dados observados que ignora o mecanismo de perda (LITTLE; RUBIN, 2002). Os métodos de imputação baseados em MAR utiliza-se de (2.3) para imputar os dados.

2.2 IMPUTAÇÃO DE DADOS

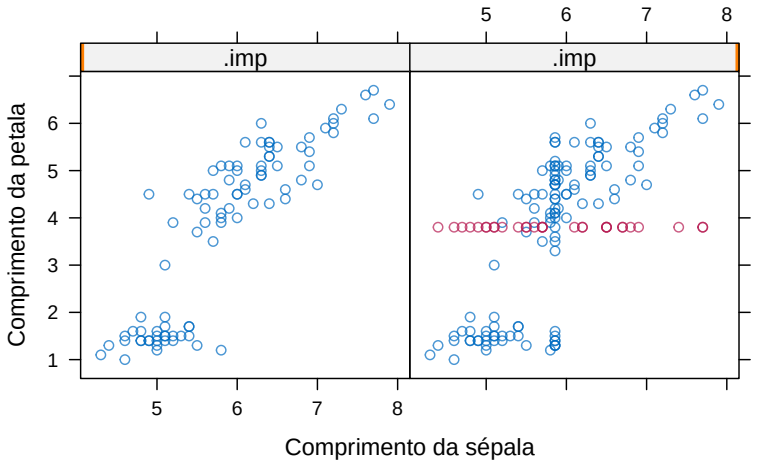
No início dos trabalhos com imputação vários mecanismos de preenchimento dos dados faltantes foram sendo propostos na literatura como substituição pela média, regressão, imputação estocástica, entre outras. Estas abordagens ad-hoc iniciais estavam baseadas em imputação única, ou seja, trabalhava-se com um único conjunto de dados preenchido por alguma técnica para se realizar toda a análise.

O método mais simples e mais comum é a substituição pela média. Para exemplificar utilizaremos o pacote *mice* do software R para imputar os dados. Este pacote permite gerar dados faltantes em um banco de dados completo pela função ampute, e faremos isto no famoso banco de dados iris onde utilizaremos o comprimento da pétala como variável resposta y e comprimento da sépala como variável preditora x. Após a geração dos valores faltantes passamos dos dados completos $(\mathbf{X}, \mathbf{Y})$ para uma estrutura na forma $(\mathbf{X}^*, \mathbf{Y}^*)$ em que $\mathbf{X}^* = (\mathbf{X}_{obs}, \mathbf{X}_{falt})$ e $\mathbf{Y}^* = (\mathbf{Y}_{obs}, \mathbf{Y}_{falt})$.

A imputação pela média é feita então pela função *mice* incluindo o método. Este tipo de imputação é a solução mais simples para dados faltantes. No entanto, a técnica pode, em

muitos casos, subestimar a variância, perturbar as relações entre as variáveis, enviesar quase qualquer estimativa diferente da média e enviesar a estimativa da média quando os dados não são MCAR (BUUREN, 2018). Na Figura 1 com as variáveis podemos ver como a imputação pela média (pontos em vermelho) não captura a variabilidade dos dados.

Figura 1 – Imputação pela média das variáveis. O vermelho indica os valores imputados.



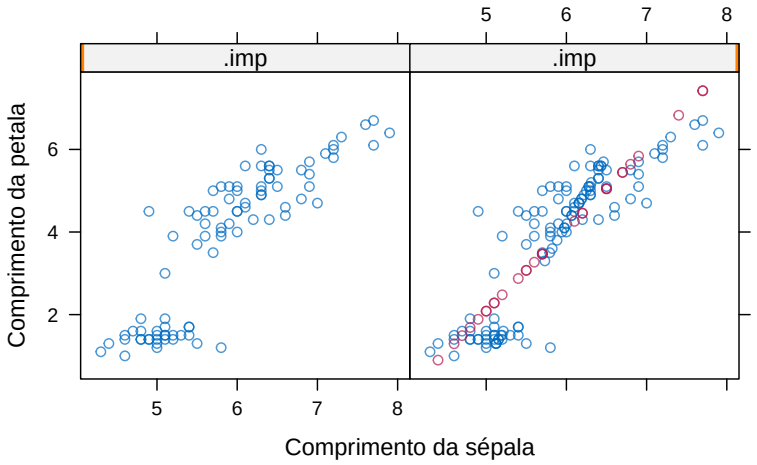
Fonte: O autor (2022).

A imputação por regressão leva em consideração o conhecimento de outras variáveis e suas relações com o intuito de produzir imputações que condizem com a realidade. Constrói-se um modelo a partir dos dados observados, e a partir das previsões calculadas para os casos incompletos de acordo com o modelo ajustado substituem-se os dados ausentes (BUUREN, 2018). A imputação por regressão simples utilizando o *mice* segue apenas mudando o parâmetro *method* da função.

Os valores imputados pelo método da regressão são os mais prováveis no modelo. No entanto, os valores imputados não capturam a variabilidade dos dados, isto é visto claramente na Figura 2 e é quase improvável que capturem a real distribuição do comprimento da pétala.

A imputação de regressão produz estimativas imparciais das médias sob MCAR, assim como a imputação da média, e dos pesos de regressão do modelo de imputação se as variáveis explicativas forem completas. Geralmente as correlações são enviesadas para cima e a variabilidade dos dados imputados é sistematicamente subestimada. O grau de subestimação depende da variância explicada e da proporção de casos ausentes (LITTLE; RUBIN, 2002).

Figura 2 – Imputação por regressão. O vermelho indica os valores imputados.



Fonte: O autor (2022).

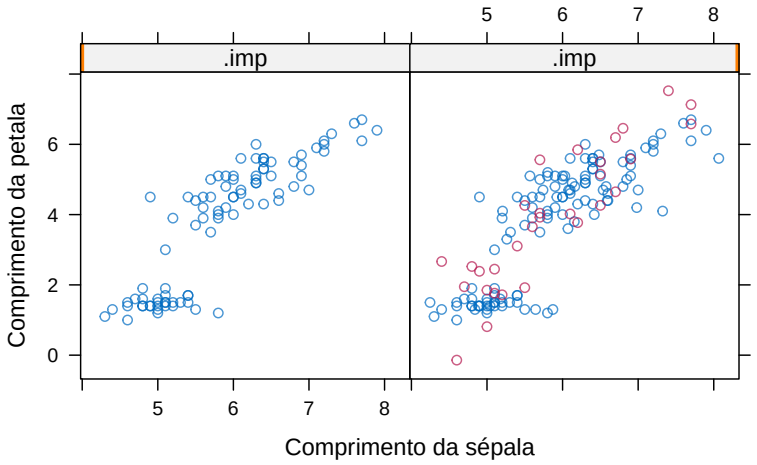
A imputação por regressão estocástica é uma tentativa de melhorar a imputação por regressão adicionando ruído às previsões para correção de viés de correlação. Este método primeiro estima o intercepto, inclinação e variância residual no modelo linear, então calcula o valor predito para cada valor ausente e adiciona um erro aleatório com base na variância estimada. A adição de ruído às previsões abre a distribuição dos valores imputados como podemos ver na Figura 3 conforme pretendido. Este método é um avanço considerável embora possa parecer contra-intuitivo não escolher a melhor previsão adicionando ruído aleatório, mas isso é precisamente o que o torna adequado para imputação.

LITTLE; RUBIN (2019) argumentam que em muitas situações estas abordagens podem não ser satisfatórias além de serem muito restritivas pois trabalhar com um único banco de dados imputado pode gerar conclusões diferentes a medida que se tem conjuntos imputados diferentes e que seria necessário uma abordagem que de fato capturasse a informação dos valores subjacentes não observados de forma consistente, a fim de evitar problemas e inconsistências nas conclusões tiradas de tais dados.

2.2.1 Imputação Múltipla em padrões de falta monotônicos e não monotônicos

O método proposto para capturar a consistência desejada nas inferências que não se têm na imputação única é a imputação múltipla (IM), que nada mais é do que a criação de vários conjuntos de dados imputados digamos D. Após as D imputações terem sido obtidas

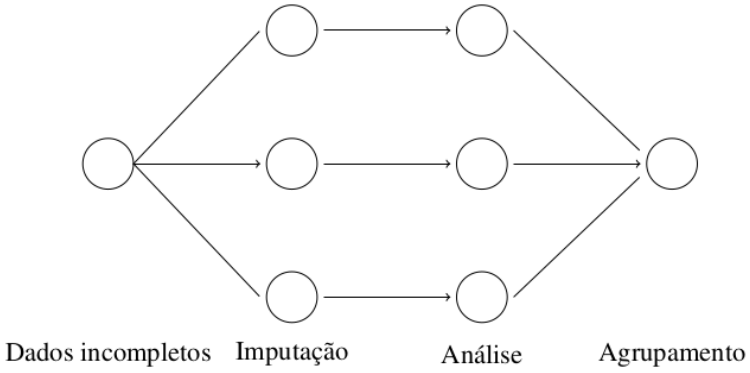
Figura 3 – Imputação por regressão estocástica. O vermelho indica os valores imputados.



Fonte: O autor (2022).

no primeiro passo, cada um dos D bancos de dados completados pela IM são analisados separadamente por métodos estatísticos tradicionais. Os D resultados são agrupados em uma estimativa de pontual final mais o erro padrão por meio de regras de agrupamento conhecidas como Regras de Rubin (RUBIN, 1981). As três etapas principais da imputação múltipla: imputação, análise e agrupamento, são representadas pelo seguinte esquema

Figura 4 – Etapas da imputação múltipla



Fonte: O autor (2022).

RUBIN (1988) descreve o conjunto de regras como se segue: em cada conjunto imputado D obtêm-se estimativas para um parâmetro de interesse θ, ou seja, θ_d, $d = 1, \ldots, D$. A estimativa geral será a média das estimativas individuais

$$\bar{\theta} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d. \tag{2.4}$$

Para a variância combinada primeiro obtêm-se a variância média dentro das imputações

$$\bar{\sigma}^2 = \frac{1}{D}\sum_{d=1}^{D}\hat{\sigma}_d^2, \tag{2.5}$$

em que σ_d^2 são as variâncias dos estimadores dentro de cada uma das imputações para $d = 1, \ldots, D$. A variância entre imputações é dada por

$$\delta = \frac{1}{(D-1)}\sum_{d=1}^{D}(\hat{\theta}_d - \bar{\theta})^2. \tag{2.6}$$

Desta forma, a variância combinada (variância total) é dada por

$$\Sigma = \bar{\sigma}^2 + \left(1 + \frac{1}{D}\right)\delta. \tag{2.7}$$

Na IM, o mais importante é a decisão na primeira etapa, ou seja, a escolha do método de IM que será utilizado para gerar as D imputações diferentes, pois é preciso que se avalie a relação das observações faltantes com as observações presentes como descrito na Seção 2.1, além de considerar o mecanismo de ausência e o padrão dos dados faltantes condizentes com a o critério de falta subjacente.

2.2.1.1 Estrutura Bayesiana em IM

Na estatística Bayesiana as inferência são realizadas em cima da distribuição posteriori. Desta forma, sob a condição de ignorabilidade, temos que a distribuição posteriori dos dados observados é dada por

$$\pi(\theta|\mathbf{Y}_{obs}) = \frac{\pi(\theta)f(\mathbf{Y}_{obs}|\theta)}{f(\mathbf{Y}_{obs})},$$

em que $\pi(\theta)$ é a distribuição priori dos parâmetros θ. Considerando a distribuição marginal $f(\mathbf{Y}_{obs})$ como uma constante normalizadora que não afeta os parâmetros de forma e escala, reduzimos a priori a

$$\pi(\theta|\mathbf{Y}_{obs}) \propto \pi(\theta)f(\mathbf{Y}_{obs}|\theta).$$

Pode-se atualizar a informação que temos sobre os dados por meio da verossimilhança dos dados observados, visto que a distribuição preditiva posteriori dos $\mathbf{Y}_{falta}$ faltantes é dada

por

$$f(\mathbf{Y}_{falt}|\mathbf{Y}_{obs}) = \int f(\mathbf{Y}_{falt}|\mathbf{Y}_{obs},\theta)\pi(\theta|\mathbf{Y}_{obs})d\theta.$$

Quando temos padrão de falta de dados monótono podemos realizar um sorteio aleatório da distribuição posteriori dos dados observados e da distribuição preditiva condicional de $\mathbf{Y}_{falta}$ dado $\mathbf{Y}_{obs}$, pois não conseguimos extrair diretamente da distribuição preditiva posterior dos dados ausentes dado os dados observados. Se o padrão de dados faltantes é não monótono então recorremos a técnicas de como técnicas de Markov Chain Monte Carlo (MCMC). Estes padrões serão definidos nas próximas seções.

2.2.1.2 Padrão de falta monotônico

Quando os dados são dispostos em forma de matriz, onde as linhas são os indivíduos e as colunas são variáveis é possível identificar padrões de não resposta. Um padrão monotônico de não-resposta é indicado na Tabela 1 (i) onde temos dados faltantes em somente uma das variáveis configurando-se o padrão univariado (caso particular do padrão monotônico) e na Tabela 1 (ii) temos um padrão onde podemos ordenar de forma que $n_{falt}(Y_{(1)}) \geq n_{falt}(Y_{(2)}) \geq \cdots \geq n_{falta}(Y_{(p)})$, ou seja, para Y_{j-1}, seus casos observados são sempre um subconjunto dos casos observados de Y_j (RUBIN, 1987).

Tabela 1 – Padrões monotônicos de falta de dados

(i)

Y_1	Y_2	$\ldots$	Y_p
y_{11}	y_{12}	$\cdots$	y_{1p}
y_{21}	y_{22}	$\cdots$	y_{2p}
-	y_{32}	$\cdots$	y_{3p}
-	y_{42}	$\cdots$	y_{4p}
-	y_{52}	$\cdots$	y_{5p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
-	y_{n2}	$\cdots$	y_{np}

(ii)

Y_1	Y_2	Y_3	$\ldots$	Y_p
y_{11}	y_{12}	y_{13}	$\cdots$	y_{1p}
y_{21}	y_{22}	y_{23}	$\cdots$	y_{2p}
-	y_{32}	y_{33}	$\cdots$	y_{3p}
-	-	y_{43}	$\cdots$	y_{4p}
-	-	-	$\cdots$	y_{5p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-	-	-	$\cdots$	y_{np}

Fonte: O autor (2022).

Seja $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{falt})$, e sejam θ os parâmetros pertencentes à verossimilhança de dados completos. A função de log-verossimilhança de dados observados pode então ser expressa como

$$l(\theta; \mathbf{Y}_{obs}) = \sum_{i=1}^{n} \ln f(\mathbf{Y}_{obs,i}|\theta). \tag{2.8}$$

O padrão monotônico é um caso especial onde a verossimilhança (2.8) pode ser fatorada em partes, de modo que a parte mais completa é condicionada às variáveis completamente observadas de $\mathbf{Y}_{obs}$, a segunda parte mais completa é condicionada as variáveis completamente observadas e a parte mais completa, e assim por diante, até que a parte menos completa seja condicionada a todas as outras partes de $\mathbf{Y}_{obs}$. A vantagem de expressar as verossimilhanças como condicionais às partes mais observadas é que cada parte pode ser maximizada separadamente, o que torna a derivação da verossimilhança dos dados observados muito mais tratável (LITTLE; RUBIN, 2019).

Dois métodos são os principais em abordagens com dados de padrão monotônico que é a BLR (Bayesian Linear Regression) e a PMM (Prediction Mean Matching).

2.2.1.3 Regressão linear bayesiana

O modelo de regressão linear é um dos métodos mais comuns de predizer Y_i em função de preditores X_i. Neste modelo, assume-se que a distribuição dos erros e_i condicionada a X_i segue uma distribuição normal com média zero e variância σ^2, o que é o mesmo que assumir que a distribuição condicional de Y dado X é normal como se segue

$$Y|\mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2),$$

em que β é um vetor de p componentes sendo p o número de preditores e σ é um escalar. Assume-se a distribuição priori não informativa $P(\beta)$ para β, e para evitar complexidade assume-se que $n_1 > p$ onde $n_!$ é o número de observações presentes.

RUBIN (1987) mostra que a distribuição a posteriori de β envolve apenas os valores Y_i efetivamente observados. Tendo a distribuição a posteriori de β descrita em termos da distribuição padrão das quais é possível retirar os dados facilmente, a estimação dos parâmetros usados na imputação está completa. (RUBIN, 1987) prova que para o modelo normal a posteriori, σ^2 é $\hat{\sigma}_1^2(n_{obs} - p)$ sobre uma variável aleatória $\chi^2_{n_1-p}$ e $\beta \sim N(\hat{\beta}_1, \sigma^2 V)$, que em termos de mínimos quadrados tem-se

$$\hat{\sigma}_1^2 = \frac{\sum_{obs}(Y_i - X_i\hat{\beta}_1)^2}{(n_{obs} - p)},$$

e

$$\hat{\beta}_1 = V\left[\sum_{obs} X_i^{\intercal} X_i\right],$$

em que $V = [\sum_{obs} X_i^{\intercal} X_i]^{-1}$. A imputação para esse modelo pode ser descrita nos seguintes passos (NUNES, 2007):

1. Simular uma variável aleatória g de uma distribuição $\chi^2_{n_{obs}-q}$, e seja

$$\sigma^{2*} = \frac{\sigma_1^2(n_{obs} - p)}{g};$$

2. Simular q variáveis aleatórias independentes $N(0, 1)$ para criar um vetor Z de p componentes tendo

$$\beta^* = \hat{\beta}_1 + \sigma^*[V]^{1/2} Z,$$

em que $V = [\sum_{obs} X_i^{\intercal} X_i]^{-1}$. A matriz $[V]^{1/2}$ pode ser vista como a raiz quadrada triangular obtida pela fatoração de Cholesky de uma matriz hermitiana (produto de uma matriz triangular inferior com entradas diagonais positivas e reais pela sua conjulgada transposta). Esta fatoração é útil por exemplo para soluções numéricas eficientes e simulações de Monte Carlo;

3. Simular os n_0 valores de Y_{falt} (valores faltantes) como

$$Y_i^* = X_i \beta^* + z_i \sigma^*,$$

em que os n_0 desvios normais z_i são simulados independentemente. Desta forma, se D imputações são desejadas, estes três passos são repetidos D vezes.

2.2.1.4 Correspondência média preditiva

A imputação por correspondência média preditiva (*prediction mean matching*) foi proposta por RUBIN (1986) e LITTLE (1988), e teve o intuito de reduzir o viés introduzido pela imputação por meio da substituição de um valor ausente por um valor real dos valores observados, o que produz uma imputação mais parecida com os valores reais observados pois estes permanecem na mesma escala. Originalmente o PMM foi formulado para ser usado em situações em que uma única variável tivesse dados ausentes ou, de forma mais ampla, quando o padrão de dados ausentes fosse monotônico (ALLISON, 2015).

Suponha que temos apenas uma variável com dados faltantes X^* e um conjunto de variáveis $\mathbf{X}$ sem dados faltantes que serão usados para imputar X. O algoritmo de PMM segue como (ALLISON, 2015):

1. Para o caso com dados faltantes, ajuste um modelo de regressão linear de X^* sobre $\mathbf{X}$, produzindo um conjunto de coeficientes β;

2. Faça um sorteio aleatório da "distribuição posteriori preditiva" de β , produzindo um novo conjunto de coeficientes β^*. Normalmente, sorteia-se aleatoriamente de uma distribuição normal multivariada com média β e a matriz de covariância estimada de β (com um sorteio aleatório adicional para a variância residual). Esta etapa é necessária para produzir variabilidade suficiente nos valores imputados;

3. Através de β^*, gere valores previstos para X^* para todos os casos, tanto aqueles com dados ausentes em X^* quanto aqueles com dados presentes;

4. Para cada caso $x_{falt,i}$ ausente, identifique um conjunto de casos, $\mathbf{x}_{obs}$ observados cujos valores previstos $\hat{x}_{obs,i}$ estejam próximos do valor previsto para o caso com dado ausente e defina $D_{i,j} = |\hat{x}_i - \hat{x}_j|$;

5. Impute $x_{obs,j}$ em $x_{falt,i}$ se $D_{i,j} \leq D_{i,k}$, $k = 1, \ldots, n_{obs}$;

6. Repita as etapas 2 a 5 para cada conjunto de dados concluído.

Ao imputar valores observados, o PMM cumpre o requisito de obter valores plausíveis. O objetivo da regressão linear não é gerar valores imputados, mas sim construir uma métrica para combinar casos com dados ausentes a casos semelhantes com dados presentes (ALLISON, 2015). Desta forma o modelo é implícito, ou seja, não é necessário definir um modelo explicito para a distribuição dos valores ausentes, o que torna a PMM menos suscetível a erros de especificação do modelo (LITTLE; RUBIN, 2002).

2.2.1.5 Padrão de falta não monotônico

Quando temos um padrão onde, por exemplo, Y_1 e Y_2 não são observados juntos como mostrado na Tabela 2 (i). As estimativas sobre esse tipo de não-resposta requer a suposição de condicionalidade de Y_1 e Y_2 dado Y_3. Já a Tabela 2 (ii) representa um padrão geral sem estrutura conhecido também como padrão arbitrário (RUBIN, 1987).

O método MCMC (Markov Chain Monte Carlo), que teve sua versão moderna proposta por Stanislaw Ulam no final dos anos 1940, é um método que pode ser utilizado para gerar uma distribuição que aproxima uma distribuição alvo f, e é indicado quando não há padrão

Tabela 2 – Padrões não monotônicos de falta de dados

(i)

Y_1	Y_2	$\ldots$	Y_p
y_{11}	-	$\ldots$	y_{1p}
y_{21}	-	$\ldots$	y_{2p}
y_{31}	-	$\ldots$	y_{3p}
y_{41}	-	$\ldots$	y_{4p}
-	y_{52}	$\ldots$	y_{5p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
-	y_{n2}	$\ldots$	y_{np}

(ii)

Y_1	Y_2	Y_3	$\ldots$	Y_p
y_{11}	-	y_{13}	$\ldots$	-
y_{21}	y_{22}	y_{23}	$\ldots$	y_{2p}
-	y_{32}	-	$\ldots$	y_{3p}
-	-	y_{43}	$\ldots$	-
y_{51}	y_{52}	-	$\ldots$	y_{5p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
-	-	-	$\ldots$	y_{np}

Fonte: O autor (2022).

monotônico porque a função de verossimilhança conjunta dos dados só pode ser fatorada em funções independentes quando o padrão da falta dos dados é monotônico.

Os métodos que são usuais quando se imputa através de MCMC são o BB (*Bayesian Bootstrap*) proposto por RUBIN (1981) e o FCS (*Fully Conditional Specifications*) especificado por SCHAFER (1997).

O Bootstrap Bayesiano (BB) é um equivalente bayesiano ao Bootstrapping clássico proposto por EFRON (1979), embora seu propósito original fosse aproximar a distribuição a posteriori de θ. Seja $\mathbf{p}^b = \{p_1^b, \ldots, p_n^b\}$ a proporção em que y_i aparece na amostra bootstrap b, $p_i^b \in 0/n, 1/n, \ldots, n/n$, com $\sum_{i=1}^n p_i^b = 1$. Podemos pensar em bootstrap como simulando esses $\mathbf{p}^b$ e atribuindo-os como pesos aos dados originais, por exemplo, a média da amostra $1/n \sum_{i=1}^n y_i^b$ é equivalente a $1/n \sum_{i=1}^n p_i^b y_i$. O bootstrap bayesiano (RUBIN, 1981) segue como: defina $\mathbf{d} = \{d_1, \ldots, d_K\}$ seja o número de vezes que cada valor distinto de $y_1, \ldots, y_K$ são observados, cada iteração das amostras de bootstrap bayesiano da seguinte distribuição posterior

$$\pi(\mathbf{p}|\mathbf{d}) \propto \pi(\mathbf{p}) L(\mathbf{d}|\mathbf{p}),$$

em que

$$L(\mathbf{d}|\mathbf{p}) \sim Multinomial(K, p_1, \ldots, p_K),$$
$$\pi(\mathbf{p}) \sim Dirichlet(\mathbf{0}),$$

que implica

$$\pi(\mathbf{p}|\mathbf{d}) \sim Dirichlet(\mathbf{d}).$$

O algoritmo do BB para imputação pode ter variações da versão original proposta por RUBIN (1981). Seja Y uma variável parcialmente observada, onde os casos $1, \ldots, n_{obs}$ são observados, e os casos $n - n_{obs}$ restantes estão faltando. Em seguida, o algoritmo é implementado da seguinte maneira (RUBIN, 1981; RUBIN, 1987):

1. Gerar $n_{obs}-1$ componentes aleatórios $U = (u_1, \ldots, u_{n_{obs}-1})$ em $[0, 1]$ da distribuição uniforme e classificar os componentes em ordem decrescente gerando $U^* = (u_{(1)}, \ldots, u_{(n_{obs-1})})$ em que $u_{(1)} < u_{(2)}, \ldots, u_{(n_{obs}-2)} < u_{(n_{obs}-1)}$. Cria-se dois vetores $n \times 1$ $W = [U^*, 1]^\intercal$ e $V = [0, U^*]^\intercal$, e calcula-se as diferenças $D = [(d_1 = w_1 - v_1), \ldots, (d_{n_{obs}} = w_{n_{obs}} - v_{n_{obs}})]^\intercal$, em que $\sum_{i=1}^{n_{obs}} d_i = 1$.

2. Realiza-se um sorteio aleatório de tamanho n_{obs} de uma distribuição multinomial com os $n_{obs} \times 1$ vetor D como pesos de probabilidade para obter n^*_{obs} da amostra.

Baseado neste conjunto de dados, sorteios aleatórios para θ são realizados. Este procedimento substitui sorteios aleatórios de uma distribuição teórica para $\theta|Y$.

Uma variante do BB, chamada Bootstrap bayesiano aproximado (BBA), foi proposta por RUBIN; SCHENKER (1986) como uma técnica de IM alternativa. Suponha que temos uma variável Y parcialmente observada e outras variáveis (completamente observadas) do mesmo conjunto de dados que podem ser usadas para estratificar a amostra em células G, com $g = 1, \ldots, G$, onde os elementos em cada célula são independentes e distribuídos de forma idêntica. $n_{obs,g}(n_{falt,g})$ denota o número de unidades observadas (ausentes) da célula g. Para cada célula, as três etapas a seguir são realizadas:

1. Obtenha uma amostra aleatória de n^*_{obs} de tamanho n_{obs} (com reposição) de n_{obs}.

2. Obtenha outra amostra aleatória de $n^*_{falt,g}$ de tamanho $n_{falt,g}$ (com reposição) de $n^*_{obs,g}$.

3. Imputa-se os valores faltantes de Y em g com valores obtidos de $n^*_{falt,g}$.

Múltiplas imputações são criadas repetindo essas etapas M vezes.

No caso onde se trabalha com as especificações totalmente condicionais o foco não estar em estimar os parâmetros da distribuição $f(\mathbf{Y}|\theta)$, mas sim modelar um subconjunto por meio de distribuições condicionais. Modelos de regressão de equações encadeadas (não restritas) são realizados variável por variável, de modo que as distribuições condicionais univariadas são dadas por $f(\mathbf{Y}_j|\mathbf{Y}_{-j}, \theta_j)$, em que $\mathbf{Y}_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$. Desta forma, em

vez de ter que lidar com um problema p-dimensional, o FCS divide a tarefa em p problemas unidimensionais (BUUREN et al., 2006).

Um algoritmo de IM que usa equações encadeadas utilizando-se de regressões passo a passo é a correspondência média preditiva bootstrap bayesiano conhecido como BBPMM e tem se tornado bem popular para imputar valores faltantes em grandes bases de dados (KOLLER-MEINFELDER, 2009). Considere que estamos com uma base dados com algumas variáveis com valores ausentes, primeiramente as variáveis são classificadas em ordem crescente de acordo com sua porcentagem de valores ausentes. Uma solução inicial é obtida regredindo cada variável com valores perdidos apenas nas variáveis completas. Desta forma, defina X_{obs}^{k} como sendo a linha com valores observados da variável resposta Y_k, com $k = 1, \ldots, K$. Temos então

$$\hat{\beta}_k = (\mathbf{X}_{obs}^{(k)\intercal}\mathbf{X}_{obs}^{(k)})^{-1}\mathbf{X}_{obs}^{(k)}Y_{k,obs},$$

e as estimativas para os valores ausentes da variável Y_k são obtidas na forma

$$\hat{\mathbf{y}}_{k,falt} = \mathbf{X}_{falt}^{(k)}\hat{\beta}_k.$$

Adicionalmente gera-se imputações por meio do PMM. No caso de todas as variáveis terem valores ausentes, a solução inicial é gerada usando sorteios aleatórios para imputar os valores ausentes na variável Y_1, antes que as variáveis Y_2 a Y_K sejam regredidas em Y_1. Um dos problemas das abordagens de equação em cadeia é que não sabemos se existe uma distribuição conjunta sobre todas as variáveis, desta forma, classifica-se as variáveis no conjunto de dados por seus respectivos números de valores perdidos. Idealmente, isso criaria um padrão de omissão monótono, para o qual existe a distribuição conjunta. Obviamente, é mais provável que o padrão após a classificação não seja monótono, mas a rotina de classificação pelo menos se aproxima do padrão monótono (KOLLER-MEINFELDER, 2009).

De posse da solução inicial, Y_k é regredido novamente nas variáveis completamente observadas e nas variáveis parcialmente imputadas Y_2 a Y_K. Na iteração r baseia-se Y_1 tanto partes completamente observadas de Y_2 a Y_K, como nas imputações destas geradas nas $r-1$ iterações.

Seja $\mathbf{X}^{(k)}$ a matriz $n \times p$ da variável regressora Y_k, então

$$\hat{\beta}_k = (\mathbf{X}^{(k)\intercal}\mathbf{X}^{(k)})^{-1}\mathbf{X}^{(k)}Y_k, \tag{2.9}$$

em que os preditores da variável Y_k são obtidos por

$$\hat{y}_k = \mathbf{X}^{(k)}\hat{\beta}_k. \tag{2.10}$$

A etapa de correspondência de média preditiva é aplicada logo após a obtenção de preditores para uma determinada variável com valores ausentes, em vez de no final de um ciclo coletivamente (KOLLER-MEINFELDER, 2009). O BBPMM segue da forma: gera-se uma amostra $x_1^*, \ldots, x_n^*$ a partir dos dados originais $X = (x_1, \ldots, x_n)$ por meio do Bootstrap Bayesiano descrito anteriormente, e esta amostra é usada para obter $\hat{\theta}^* = \hat{\theta}(x_1^*, \ldots, x_n^*)$ que está substituindo os sorteios de distribuição priori. Assim, as equações (2.9) e (2.10) são posteriormente substituídas por

$$\hat{\beta}_k^* = (\mathbf{X}^{(k),*\intercal}\mathbf{X}^{(k),*})^{-1}\mathbf{X}^{(k),*}Y_k^*, \tag{2.11}$$

e

$$\hat{y}_k = \mathbf{X}^{(k)}\hat{\beta}_k^*. \tag{2.12}$$

O i-ésimo passo também é substituído pelo PMM, conforme descrito acima. Este é o algoritmo BBPMM que funciona bem em grandes conjuntos de dados com padrões de dados ausentes não monótonos.

3 MODELOS DE REGULARIZAÇÃO

3.1 REGRESSÃO LOGÍSTICA PENALIZADA

No contexto de regressão o interesse é modelar como uma característica populacional a qual denomina-se variável resposta (dependente) é afetada por uma ou mais variáveis denominadas explicativa (covariável, variável independente). Deseja-se modelar o comportamento da variável resposta denotada por y levando em consideração o impacto que este comportamento sofre em relação a variável explicativa denotada por x, ou seja, uma regressão de y sobre x é qualquer aspecto da distribuição de y condicional a x. Mais especificamente, busca-se estimar a função de regressão

$$f(\mathbf{x}) = \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}), \tag{3.1}$$

em que $\mathbf{X}$ é uma matriz $n \times p$ (p número de covariáveis, n tamanho da amostra) e $\mathbf{Y}$ o vetor $n \times 1$ de variáveis resposta. Métodos paramétricos assumem que a função de regressão $f(\mathbf{x})$ pode ser parametrizada com um número finito de parâmetros. Quando a variável resposta é binária $Y \in \{0, 1\}$, o modelo logístico linear é frequentemente usado: ele modela a razão de probabilidade logarítmica como a combinação linear

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \beta_0 + \mathbf{X}^\intercal \beta, \tag{3.2}$$

em que $\mathbf{X} = (X_1, \ldots, X_p)$ é o vetor de covariáveis, $\beta_0 \in \mathbb{R}$ é o intercepto e $\beta \in \mathbb{R}^p$ é o vetor de coeficientes de regressão. Invertendo esta transformação obtem-se a probabilidade condicional

$$P(Y = 1|X = x) = \frac{\exp(\beta_0 + \mathbf{X}^\intercal \beta)}{1 + \exp(\beta_0 + \mathbf{X}^\intercal \beta)}. \tag{3.3}$$

Normalmente, ajustamos modelos logísticos maximizando a probabilidade logarítmica binomial dos dados. Desta forma, a função de log-verossimilhança para estimação dos parâmetros é dada por

$$l(\beta_0, \beta) = \sum_{i=1}^{n} \left[y_i(\beta_0 + x_i^\intercal \beta) - \log\left(1 + \exp(\beta_0 + x_i^\intercal \beta)\right) \right]. \tag{3.4}$$

No contexto de aprendizado de máquina o objetivo é utilizar o modelo para predição de novas observações, e deseja-se o modelo com menor erro de predição. Quando há muitas covariáveis $p > n$, podemos sofrer com baixo desempenho preditivo devido ao super-ajuste

(bom ajuste para os dados em questão porém com péssimo desempenho preditivo para dados futuros), ou na presença de multicolinearidade (alta correlação entre as covariáveis) há problemas com relação a solução de (3.4).

Uma forma de contornar estes problemas é a regressão regularizada que consiste em adicionar um termo de penalização no conjunto de funções para diminuir os espaço de busca dos β_j $j = 1, \ldots, p$, na busca de diminuir a complexidade da regressão para diminuição da variância da função de predição estimada buscando pelo melhor subconjunto de covariáveis, ou seja, essa solução consiste em aumentar o viés do estimador de (3.1) na busca de uma diminuição significativa de sua variância (IZBICKI; SANTOS, 2020). Muitas vezes as penalizações pioram a predição nos dados de treino (onde se ajusta o modelo) para poder melhorar a predição de teste (validação do modelo), evitando o super-ajuste. Os modelos mais comuns de regularização são o Rigde, LASSO e Rede Elástica e serão apresentados a seguir.

3.1.1 Penalização Rigde

O modelo de regressão Rigde proposto por HOERL; KENNARD (1970) restringe o espaço de busca dos β's com a penalização conhecida como L_2. A família de distribuições do conjunto restrito para a abordagem Rigde é dada por

$$\mathcal{F} = \left\{ f(\mathbf{x}) = \beta_0 + \mathbf{X}^\intercal \beta, \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, ||\beta||_2^2 \leq B \right\}, \tag{3.5}$$

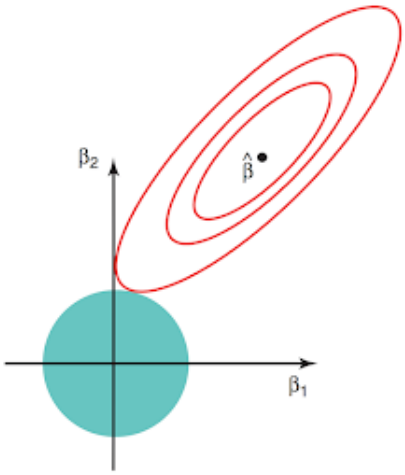
em que a penalização $||\beta||_2^2 = \sum_{j=1}^p \beta_j^2$. Desta forma, escolher a melhor função do ponto de vista preditivo é o mesmo que minimizar o erro (a prova desta afirmação será omitida pois foge aos objetivos deste trabalho), que no caso do modelo de regressão logística buscamos minimizar

$$\hat{E}_{(XY)} = \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \left[y_i(\beta_0 + x_i^\intercal \beta) - \log\left(1 + \exp(\beta_0 + x_i^\intercal \beta)\right) \right] + \lambda ||\beta||_2^2 \right\}, \tag{3.6}$$

em que $\lambda > 0$ é o parâmetro que regulariza a contribuição da penalização, geralmente $\lambda \in (0, 1)$. Apesar de (3.6) não fazer seleção de variáveis como o LASSO que veremos a seguir, a regressão Ridge também diminui a variância dos estimadores da regressão pois encolhe os coeficientes β's estimados pela regressão linear restringindo o espaço de busca.

A Figura 5 é uma representação da regressão Rigde (3.5) para o modelo linear no caso com duas covariáveis em que a área azul representa a região de restrição, o mapa de contorno em vermelho representa função de mínimos quadrados e o ponto de contato entre a elipse e a área azul é o coeficiente da regressão regularizada.

Figura 5 – Esquema gráfico da regressão Rigde no caso em que existem apenas duas covariáveis.



Fonte: (HASTIE; TIBSHIRANI; WAINWRIGHT, 2019).

Evidentemente, apesar da variância da regressão Ridge ser menor, seu viés é maior. Assim, λ deve ser escolhido de modo a controlar o balanço viés-variância. Isto pode ser feito via validação cruzada como será visto na seção 3.2.

3.1.2 Penalização LASSO

O LASSO, desenvolvido por TIBSHIRANI (1996) é semelhante a proposta anterior com uma mudança na penalização utilizando $\sum_{j=1}^{p} |\beta_j|$ conhecida como norma L_1. Uma vantagem da penalização L_1 é que ela captura a ideia de que uma pequena mudança nos valores de β não altera demasiadamente a complexidade do modelo resultante (IZBICKI; SANTOS, 2020). A família de distribuições do conjunto restrito para o LASSO é dada por

$$\mathcal{F} = \{f(\mathbf{x}) = \beta_0 + \mathbf{X}^\intercal\beta, \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, ||\beta||_1 \leq B\}\,, \tag{3.7}$$

em que $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$. Novamente, do ponto de vista preditivo podemos olhar na perspectiva de minimização do erro, que no caso do modelo de regressão logística buscamos minimizar

$$\hat{E}_{(XY)} = \underset{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \left[y_i(\beta_0 + x_i^\intercal\beta) - \log\left(1 + \exp(\beta_0 + x_i^\intercal\beta)\right)\right] + \lambda||\beta||_1 \right\}, \tag{3.8}$$

Note que valores diferentes de λ leva a diferentes conjuntos estimados $\hat{\beta}$ e quando $\lambda \to \infty$ o estimador é tal que $\hat{\beta}_1 = 0, \ldots, \hat{\beta}_p = 0$, ou seja, temos o modelo apenas com intercepto e o estimador dado pelo LASSO tem variância próxima a zero, mas um viés muito alto.

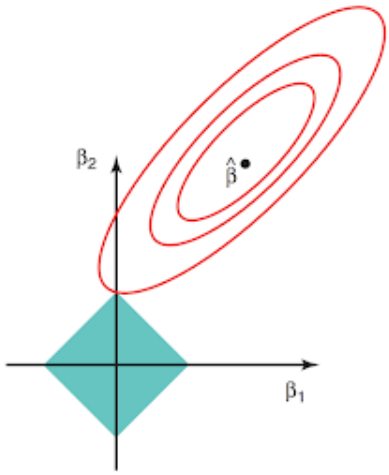
Um problema com o LASSO é que ele penaliza todos os coeficientes igualmente, através de λ. Nós esperaríamos que algumas variáveis sejam mais importantes que outras e isso pode vir a priori ou ser dito pelos dados. Uma alternativa é o aLASSO (*adaptative lasso*), em que as penalidades sofrem influência de um termo de adaptação (pesos) em que a única exigência desses pesos é que eles sejam positivos. Estes pesos adaptativos são conhecidos por abordar questões de consistência de estimativa e seleção. A função de penalização do aLASSO é dada por

$$\sum_{j=1}^{p} \hat{a}_j |\beta_j|,$$

em que $\hat{a}_j = (\hat{\beta}_j^0 + 1/n)^\gamma$, com $\gamma > 0$, $\gamma = \lceil 2v/1 - v \rceil + 1$, em que $v = \log(p)/\log(n)$ e $\hat{\beta}_j^0$ é uma estimativa inicial de $\hat{\beta}_j$ geralmente determinada por mínimos quadrados ordinários (MQO) ou estimativa de máxima verossimilhança quando $p < n$.

A Figura 6 é uma representação do LASSO para o modelo linear com duas covariáveis em que a área azul representa a região de restrição, o mapa de contorno em vermelho representa a função de mínimos quadrados e o ponto de contato entre a elipse e a área azul é o coeficiente da regressão regularizada.

Figura 6 – Esquema gráfico do LASSO no caso em que existem apenas duas covariáveis.



Fonte: (HASTIE; TIBSHIRANI; WAINWRIGHT, 2019).

A estimativa sem a penalização seria no ponto $\hat{\beta}$, porém a penalização na formulação alternativa restringe uma região e no LASSO os coeficientes podem zerar (nisto o LASSO faz seleção de variáveis) diferentemente do Rigde, note que o contorno toca o eixo β_2.

3.1.3 Penalização Elastic Net

O método de rede elástica é uma junção convexa dos dois métodos apresentados anteriormente Rigde e LASSO. A família de distribuições do conjunto restrito para o Elastic Net é dada por

$$\mathcal{F} = \left\{ f(\mathbf{x}) = \beta_0 + \mathbf{X}^\intercal \beta, \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \alpha||\beta||_1 + (1-\alpha)||\beta||_2^2 \leq B \right\}, \tag{3.9}$$

em que $\alpha \in (0,1)$, onde para $\alpha = 1$ se resume ao LASSO, e $\alpha = 0$ se resume ao Rigde. Novamente, olhando este problema pela perspectiva de minimização do erro dentro da amostra temos no caso da regressão logística temos

$$\hat{E}_{(XY)} = \underset{\beta \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \left\{ -\frac{1}{n} l(\beta_0, \beta) + \lambda(\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2) \right\}. \tag{3.10}$$

em que $l(\beta_0, \beta)$ é dada por (3.4). Desta forma, a regularização de rede elástica é algo entre a Rigde e o Lasso, sendo que graficamente sua função de restrição para o caso do modelo linear se aproxima do losango Figura 6 (LASSO) a medida que α se aproxima de 1 e próxima de um círculo Figura 5 (Ridge) a medida que α se aproxima de 0. Assim como no LASSO, a rede elástica também faz seleção de variáveis.

3.1.4 Estimação dos parâmetros

A log-verossimilhança dada em (3.4) é uma função côncava dos parâmetros. As estimativas dadas por esta função coincidem com as estimativas obtidas por mínimos quadrados ponderados. Portanto, se as estimativas atuais dos parâmetros são $\tilde{\beta}_0$,$\tilde{\beta}$, formamos uma aproximação quadrática para log-verossimilhança através de uma expansão de Taylor sobre as estimativas atuais dada por

$$l_Q(\beta_0, \beta) = -\frac{1}{2n} \sum_{i=1}^{n} w_i (z_i - \beta_0 - x_i^\intercal \beta) + c(\tilde{\beta}_0, \tilde{\beta})^2, \tag{3.11}$$

em que

$$\begin{aligned} w_i &= \tilde{p}(x_i)(1 - \tilde{p}(x_i)), \\ z_i &= \tilde{\beta}_0 + x_i^\intercal \tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{p}(x_i))}. \end{aligned}$$

e $\tilde{p}(x_i)$ é avaliado nos parâmetros atuais. O último termo é constante, e a atualização é obtida minimizando l_Q. Para cada valor de λ, criamos um loop externo que calcula a aproximação

quadrática sobre os parâmetros atuais $\tilde{\beta}_0$ e $\tilde{\beta}$. Segundo HASTIE; TIBSHIRANI; WAINWRIGHT (2019) o algoritmo de coordenada descendente é o melhor método para resolver o problema dos mínimos quadrados ponderados

$$\underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \left\{ -l_Q(\beta_0, \beta) + \lambda P_\alpha(\beta) \right\}, \tag{3.12}$$

pois a função objetivo (3.12) é convexa e a parte de verossimilhança é diferenciável, o que torna encontrar a solução uma tarefa padrão de otimização convexa. Um exemplo de algoritmo comum para encontrar uma solução para esta função é o método iterativo *proximal-Newton* que aproxima repetidamente a log-verossimilhança negativa por uma função quadrática. A penalização $P_\alpha(\beta)$ é escolhida pelo pesquisador podendo ser as de Ridge, LASSO ou rede elástica.

3.2 VALIDAÇÃO CRUZADA ANINHADA

É notória a importância do parâmetro λ na penalização tendo um papel central na regressão regularizada. Escolhe-lo corretamente é extremamente importante. O método mais comum para a escolha do λ é a validação cruzada (*cross validation*).

O método consiste em criar novas amostras artificiais a partir do conjunto de treinamento dos dados (base de dados original separada em treinamento e teste) para, através de divisões aleatórias, fazer a estimação da performance do modelo para diferentes valores de λ (geralmente $\lambda \in [0, 1]$) em cada uma das amostras buscando encontrar o λ que maximiza a performance do modelo.

Mais especificamente, divide-se o conjunto de treinamento em $K > 1$ subconjuntos (dobras) em que um destes é fixado para teste (para validação do modelo) e os demais $K - 1$ são usados para estimação do modelo para diferentes valores de λ. O conjunto fixado k é utilizado para avaliar a performance do modelo por uma função de perda $l(\hat{y}, y)$, por exemplo, a perda quadrática $(f(x) - y_i)^2$ para variáveis contínuas ou a função $\{0, 1\}-$ a perda na classificação para variáveis binarias, as quais denota-se por e_i. Temos então que uma quantidade de interesse seria o erro de predição no conjunto de treinamento (X, Y) dado por

$$E_{(XY)} = \mathbb{E}\left[\, l(f(X_{n+1}, \hat{\theta}), Y_{n+1}) | (X, Y) \right]. \tag{3.13}$$

Denotamos a esperança dessa quantidade em todos os conjuntos de treinamento possí-

veis como $E = \mathbb{E}[E_{(XY)}]$. Desta forma, a estimativa de VC para $E_{(XY)}$ é dada por

$$\hat{E}^{(VC)} = \bar{e} = \frac{1}{k}\sum_{k=1}^{K}\frac{1}{m}\sum_{i\in T_k} e_i, \tag{3.14}$$

em que m é o tamanho da dobra e T_k é o k-ésima dobra utilizada para teste. Calcula-se então o erro padrão de VC e escolhe-se o λ com o erro padrão de validação cruzada mínimo. Se deseja-se construir um intervalo de confiança (IC) pro $\hat{E}^{(VC)}$, calcula-se o erro padrão empírico

$$\widehat{SE}(\hat{e}) = \frac{1}{n}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \hat{e})^2}, \tag{3.15}$$

obtendo-se um IC

$$(\bar{e} - z_{1-\alpha/2}\widehat{SE}(\bar{e}); \bar{e} + z_{1-\alpha/2}\widehat{SE}(\bar{e})), \tag{3.16}$$

em que $z_{1-\alpha/2}$ é o quantil $1-\alpha/2$ da distribuição normal padrão. Uma questão levantada por BENGIO; GRANDVALET (2004) é sobre a estrutura da variância da validação cruzada. Os autores argumentam que a variância do estimador de validação cruzada é uma combinação linear de três momentos, a variância da dobra K, a covariância dos erros dentro de uma mesma dobra e a covariâncias dos erros entre as dobras. Desta forma, a variância de $\bar{e}$ seria definida como

$$Var[\bar{e}] = \frac{1}{n^2}\sum_{i,j} Cov(e_i, e_j) = \frac{1}{n}\sigma^2 + \frac{m-1}{n}\omega + \frac{n-m}{n}\gamma, \tag{3.17}$$

em que $\forall i, Cov(e_i, e_i) = \sigma^2$, $\forall (i,j) \in T_k^2 : j \neq i, Cov(e_i, e_j) = \omega$ e $\forall (i,j) \in T_l : l \neq k, Cov(e_i, e_j) = \gamma$ sendo ω e γ as covariâncias intra e entre dobras respectivamente. A variância tem um papel crucial na construção de intervalos de confiança para o erro de predição e a estimativa proposta em (3.17) trouxe uma nova perspectiva para intervalos de confiança em estimativas de validação cruzada.

BATES; HASTIE; TIBSHIRANI (2021) argumentam que os intervalos de confiança gerados pela variância usual da estimativa de validação cruzada não cobrem de forma consistente o erro de predição por serem menores, o que os autores chamam de intervalos ingênuos. Uma forma de criar intervalos mais abrangentes e condizentes com a realidade seria utilizando a estrutura de variância proposta em (3.17). Neste caso, bastaria estimar as quantidades ω e γ, porém, em BENGIO; GRANDVALET (2004) prova-se o fato de que não há um estimador imparcial de $Var[\bar{e}]$ baseado em uma única execução de validação cruzada. Para contornar esse problema, BATES; HASTIE; TIBSHIRANI (2021) desenvolvem um estimador que estima empiricamente a variância do erro de validação cruzada em muitas subamostras.

O erro quadrático médio (EQM) da validação cruzada para uma amostra de tamanho n com K dobras é dado por

$$EQM_{K,n} = \mathbb{E}\left[\left(\hat{E}^{(VC)} - E_{(XY)}\right)^2\right]. \tag{3.18}$$

O EQM possuí duas componentes sendo um termo de viés e um termo de variância, mas o viés é tipicamente pequeno para validação cruzada (EFRON, 1986; EFRON; GONG, 1983; EFRON; TIBSHIRANI, 1997). Desta forma, podemos ver o EQM na validação cruzada como uma estimativa conservadora da variância do estimador de VC. Com isso em mente, usaremos uma estimativa do EQM para construir intervalos de confiança para $E_{(XY)}$.

(BATES; HASTIE; TIBSHIRANI, 2021) fornecem uma decomposição genérica do erro quadrático médio de uma estimativa do erro de predição que permitirá a estimação de $EQM_{K,n}$. Considere o conjunto dos dados $\mathcal{I} = 1, \ldots, n$ particionado em $\mathcal{I}_{treino} = (\tilde{X}, \tilde{Y})$ e $\mathcal{I}_{valid}$. Usamos $(\tilde{X}, \tilde{Y})$ para estimar os parâmetros $\hat{\theta}^{treino} = f(\tilde{X}, \tilde{Y})$ e ainda assumir que temos alguma estimativa $\hat{E}_{(\tilde{X}\tilde{Y})}$ do erro de predição definido em (3.13) para a amostra de treinamento $(\tilde{X}, \tilde{Y})$. Seja $\{e_i^{(valid)}\}_{i \in \mathcal{I}_{valid}}$ as perdas do modelo ajustado $\hat{f}(., \hat{\theta}^{(treino)})$ no conjunto de validação, e $\bar{e}^{(valid)}$ a sua média. O EQM de $\hat{E}_{(\tilde{X}\tilde{Y})}$ pode ser decomposto na forma

$$\mathbb{E}\left[\left(\hat{E}_{(\tilde{X}\tilde{Y})} - E_{(\tilde{X}\tilde{Y})}\right)^2\right] = \mathbb{E}\left[\left(\hat{E}_{(\tilde{X}\tilde{Y})} - \bar{e}^{(valid)}\right)^2\right] - \mathbb{E}\left[\left(\bar{e}^{(valid)} - E_{(\tilde{X}\tilde{Y})}\right)^2\right]. \tag{3.19}$$

Note que a esperança acima é sobre os dados completos (X, Y). Essa identidade é de interesse uma vez que os termos da direita podem ser obtidos diretamente dos dados o que leva a uma estimativa do EQM. O seguinte algoritmo é proposto para a obtenção do estimador (BATES; HASTIE; TIBSHIRANI, 2021):

1 Divida seus dados em conjunto de treinamento $\mathcal{I}_{treino}$ e validação $\mathcal{I}_{valid}$, e para cada divisão faça

   (i) Calcule $\hat{E}_{(\tilde{X}\tilde{Y})}$ e $\bar{e}^{(valid)}$ e então calcule o primeiro termo da expressão;

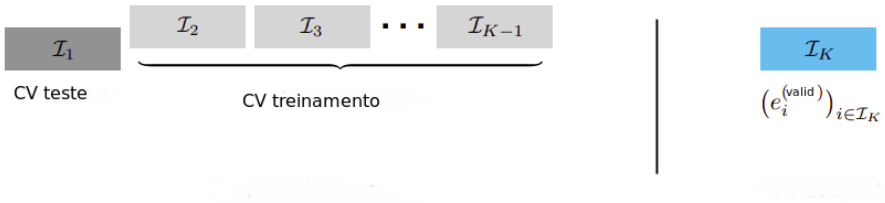
   (ii) Calcule o segundo termo da expressão através da variância empírica dos erros $\{e_i\}_{i \in \mathcal{I}_{valid}}$.

2 Faça a média das estimativas conjuntas de (i) e (ii) em todas as divisões aleatórias e tome sua diferença como em (3.19).

A validação cruzada aninhada (*nested cross validation*) é uma estratégia para o caso particular em que $\hat{E}_{(\tilde{X}\tilde{Y})}$ é em si uma estimativa de validação cruzada baseada apenas em

$(\tilde{X}, \tilde{Y})$. Segue-se a estratégia de estimação descrita acima usando $(K-1)$ dobras de VC como o estimador $\hat{E}_{(\tilde{X}\tilde{Y})}$. Também obtém-se uma estimativa pontual de erro calculando a média empírica de $\hat{E}_{(\tilde{X}\tilde{Y})}$ entre as várias divisões que é denotada por $\hat{E}^{(VCA)}$.

Figura 7 – Esquema de validação cruzada aninhada.



Fonte: (BATES; HASTIE; TIBSHIRANI, 2021).

Em vista de (3.19), o estimador $\widehat{EQM}$ tem como alvo o EQM de $\hat{E}^{(VC)}$ como uma estimativa de $E_{(XY)}$. Para um VC aninhado com uma amostra de tamanho n, temos

Teorema 3.2.1 *Para uma VCA com uma amostra de tamanho n, temos*

$$\mathbb{E}\left[\widehat{EQM}\right] = EQM_{(K-1),n^*},$$

em que $n^* = n(K-1)/K$ (BATES; HASTIE; TIBSHIRANI, 2021).

Este estimador é utilizada para construção de intervalos de confiança com uma maior cobertura e pode ser interpretada da seguinte maneira: em subamostras repetidas, estimamos a quantidade (i), primeiro termo da equação (3.19), o que leva a uma estimativa empírica de quanto um intervalo em torno de $\hat{E}_{(\tilde{X}\tilde{Y})}$ deve ser alargado a fim de cobrir $\bar{e}^{(valid)}$. A última é uma quantidade aleatória, no entanto, deve ser considerada como um intervalo de predição calibrado. Na verdade, não desejamos cobrir $\bar{e}^{(valid)}$, mas a média de $E_{(\tilde{X}\tilde{Y})}$ (BATES; HASTIE; TIBSHIRANI, 2021). Convertemos de um intervalo de predição para $\bar{e}^{(valid)}$ em um intervalo de confiança para $E_{(\tilde{X}\tilde{Y})}$ subtraindo o termo (ii), segundo termo da equação (3.19).

Para construir o intervalo de confiança é necessário fazer uma correção de viés. A estimativa de VCA do erro de predição é não viesada para E_{rr} para o procedimento com um tamanho de amostra reduzido de $n(K-2)/K$, mas é ligeiramente inclinado para cima para E_{rr} com o tamanho total da amostra. Esta discrepância pode ser estimada executando o VC usual K-dobras e o VCA (BATES; HASTIE; TIBSHIRANI, 2021). Uma estimativa imparcial para a

diferença em E_{rr} em uma amostra de tamanho $n(K-2)/K$ (o tamanho da amostra usada na VCA) para uma amostra de tamanho $n(K-1)/K$ (o tamanho da amostra usada no padrão VC) é

$$\hat{E}^{(VCA)} - \hat{E}^{(VC)}.$$

Um estimador para a diferença em E_{rr} em uma amostra de tamanho n para E_{rr} em uma amostra de tamanho $n(K-2)/K$ (o tamanho da amostra de cada modelo usado no CV aninhado) é dado na forma

$$\hat{B} := \left(1 + \left(\frac{K-2}{K}\right)\right)\left(\hat{E}^{(VCA)} - \hat{E}^{(VC)}\right). \tag{3.20}$$

O termo esquerdo na soma, "1", explica o viés ao passar do tamanho $n(K-2)/K$ para o tamanho $n(K-1)/K$ e o termo direito na soma, explica o viés ocorrendo de um tamanho de amostra de $n(K-1)/K$ ao tamanho n. Combinar isso com a estimativa de EQM na seção anterior leva aos seguintes intervalos de confiança:

$$\left(\hat{E}^{(VCA)} - \hat{B} - q_{1-\alpha/2}\sqrt{\frac{K-1}{K}}\sqrt{\widehat{EQM}}; \hat{E}^{(VCA)} - \hat{B} + q_{1-\alpha/2}\sqrt{\frac{K-1}{K}}\sqrt{\widehat{EQM}}\right). \tag{3.21}$$

A forma (3.21) produz intervalos de confiança mais largos do que os intervalos dados por (3.16), cobrindo de forma consistente a estimativa do erro de predição dado pela validação cruzada.

3.3 MODELOS DE REGULARIZAÇÃO PARA DADOS MULTI-IMPUTADOS

No contexto de dados multi-imputados modelos como de LASSO, Ridge e Rede Elástica podem trazer alguns problemas estruturais na sua implementação. Para cada conjunto imputado quando aplicado a penalização de LASSO, por exemplo, é possível que tenha-se diferentes coeficientes zerados nos vetores estimados para cada conjunto o que dificulta sua utilização juntamente com a metodologia de imputação múltipla. CHEN; WANG (2013) argumentam que considerar o ajuste do modelo em todos os conjuntos de dados imputados de forma conjunta produz uma seleção de variável consistente em todos os conjuntos de dados imputados.

DU et al. (2020) propõem uma metodologia que incorpora as duas abordagens (regularização e imputação múltipla) e definem seus métodos como se segue. Seja $\mathbf{X}_n$ a matriz $n \times p$ de covariáveis, e $\mathbf{Y}_d$ 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 $X_{d,ij}$ 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 dados é dado por $\boldsymbol{\beta}_d$, 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 $\boldsymbol{\beta}_{0,d}$. O vetor de parâmetros de regressão para o d-ésimo conjunto de dados é dado por $\boldsymbol{\theta}_d = (\boldsymbol{\beta}_{0,d}, \boldsymbol{\beta}_d)$. Uma abordagem comum na prática é usar regras ad hoc para determinar o conjunto final de variáveis selecionadas. Muitas vezes, isso prossegue ajustando um procedimento de regressão penalizado em cada conjunto de dados imputado separadamente:

$$\hat{\boldsymbol{\theta}}_d = \arg\min_{\boldsymbol{\theta}_d} \left\{ -\frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{\theta}_d | y_{d,i}, \mathbf{X}_{d,i}) + \lambda P_\alpha(\boldsymbol{\beta}_d) \right\}, \tag{3.22}$$

em que $d = 1, \ldots, D$, $P_\alpha(\boldsymbol{\beta}_d)$ é a função de penalidade de β_d parametrizada por α e $\lambda \in [0, \infty)$ é um parâmetro de ajuste que controla a contribuição relativa de a penalidade. Temos as seguintes penalidades:

- LASSO: $P_\alpha(\boldsymbol{\beta}_d) = \sum_{j=1}^{p} |\beta_{d,j}|$;
- aLASSO: $P_\alpha(\boldsymbol{\beta}_d) = \sum_{j=1}^{p} \hat{a}_{d,j} |\beta_{d,j}|$;
- ENET: $P_\alpha(\boldsymbol{\beta}_d) = \alpha \sum_{j=1}^{p} |\beta_{d,j}| + (1-\alpha) \sum_{j=1}^{p} \beta_{d,j}^2$;
- aENET: $P_\alpha(\boldsymbol{\beta}_d) = \alpha \sum_{j=1}^{p} \hat{a}_{d,j} |\beta_{d,j}| + (1-\alpha) \sum_{j=1}^{p} \beta_{d,j}^2$.

Temos que o ENET é a penalização de rede elástica que é a generalização da penalização LASSO e Ridge, estas sendo obtidas quando $\alpha = 1$ e $\alpha = 0$ respectivamente, e $\hat{a}$ o peso adaptativo. Todas estas penalidades já foram introduzidas no Capítulo 3.

3.3.1 Funções objetivo empilhada e agrupada

Em uma função objetivo agrupada homogênea (empilhada), somamos as funções objetivo para cada um dos conjuntos de dados imputados, obtendo assim uma função objetivo conjunta

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ -\frac{1}{n} \sum_{d=1}^{D} \sum_{i=1}^{n} l(\boldsymbol{\theta} | y_{d,i}, \mathbf{X}_{d,i}) + \lambda P_\alpha(\boldsymbol{\beta}) \right\}.$$

Note que θ não esta fixado pelo índice d, isto implica que ao otimizar a função objetivo combinada resultará em um vetor de parâmetros estimado $\hat{\boldsymbol{\theta}}$, assim, impondo a seleção uniforme em todos os conjuntos de dados imputados.

No entanto, DU et al. (2020) argumentam que empilhar todos os conjuntos de dados imputados pode ser visto como um aumento artificial do tamanho da amostra . Uma maneira comum sugerida para resolver isso é adicionar um peso de observação o_i, de modo que o peso total para cada sujeito no conjunto de dados empilhado soma um. Um peso alternativo é dado por $o_i = f_i/D$ em que f_i é a frequência (número) de covariáveis do total de covariáveis para o sujeito i, com isto, o peso dos sujeitos com menos dados faltantes é maior. Logo, a função objetivo pode ser reescrita da forma

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ -\frac{1}{n} \sum_{d=1}^{D} \sum_{i=1}^{n} o_i l(\boldsymbol{\theta} | y_{d,i}, \mathbf{X}_{d,i}) + \lambda P_\alpha(\boldsymbol{\beta}) \right\}. \tag{3.23}$$

Em uma função objetivo agrupada heterogênea, se impõe a seleção de variáveis uniformes em conjuntos de dados imputados adicionando um termo de grupo na penalização na função objetivo, na forma

$$(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_D) = \arg\min_{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D} \left\{ -\frac{1}{n} \sum_{d=1}^{D} \sum_{i=1}^{n} l(\boldsymbol{\theta} | y_{d,i}, \mathbf{X}_{d,i}) + \lambda P_\alpha(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D) \right\}. \tag{3.24}$$

CHEN; WANG (2013) originalmente formularam um caso especial da função objetivo agrupada heterogênea conhecida como MI-LASSO, onde a função de penalidade é dada por

$$P_\alpha(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D) = \sum_{j=1}^{p} \sqrt{\sum_{d=1}^{D} \beta_{d,j}^2}. \tag{3.25}$$

Os modelos de regularização com a penalizações (3.23) e (3.24) serão utilizados e avaliados quanto ao desempenho preditivo nas aplicações do Capítulo 5 frente ao modelo usual de LASSO para dados completos. As duas penalidades que iremos utilizar são dadas a seguir

- GaLASSO: $P(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D) = \sum_{j=1}^{p} \hat{a}_j |\beta_{d,j}|$;
- SaENET: $P_\alpha(\boldsymbol{\beta}_d) = \alpha \sum_{j=1}^{p} \hat{a}_j |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2$.

Estas são conhecidas como a rede elástica adaptativa empilhada SaENET da sigla em inglês e o LASSO de grupo adaptativo conhecido como GaLASSO.

4 ANÁLISE ROC E CURVAS DE DECISÃO

4.1 ANÁLISE ROC

Em problemas de aprendizado de máquina envolvendo classificação, os gráficos de características de operação do receptor (ROC) são úteis para visualizar o desempenho de um determinado modelo de classificação como os apresentados no Capítulo 3, comumente utilizados na tomada de decisão médica para determinar um prognóstico (FAWCETT, 2006).

Inicialmente os gráficos ROC foram usados na teoria de detecção de sinal para representar a compensação entre as taxas de acerto e as taxas de falsos alarmes dos classificadores (EGAN; EGAN, 1975), sendo posteriormente estendida para uso na visualização e análise do comportamento de sistemas de diagnóstico (SWETS, 1988).

4.1.1 Performance de Classificadores

Por simplicidade, consideremos o problema de classificação de apenas duas classes. Formalmente, cada instância $\mathbf{I}$ é mapeada para um elemento do conjunto $\{p, n\}$, classes positivo e negativo respectivamente. Alguns classificadores geram como saída valores contínuos (um classificador que gera a probabilidade de uma instância pertencer a uma dada classe), para os quais diferentes limiares podem ser aplicados para gerar diferentes conjuntos de saída (em um classificador contínuo pode-se aplicar um limiar para obter um valor binário, ou seja, gerar duas classes). Outros, geram resultados discretos indicando somente a classe (FAWCETT, 2006).

Temos então quatro situações para a classificação e as instâncias: (i) se a instância for positiva e é classificada como positiva temos um verdadeiro positivo, (ii) se for classificado como negativo temos um falso negativo, (iii) se a instância for negativa e classificada como negativa temos um verdadeiro negativo, (iv) se for contado como positivo temos um falso positivo. Esta configuração pode ser representada em uma matriz denominada matriz de confusão como é mostrada na Tabela 3, diferenciamos a classe prevista por $\{p^*, n^*\}$.

Os valores da diagonal da matriz representam as decisões corretas, e os valores fora da diagonal representam os erros (confusões) entre as classes. A partir da matriz de confusão podemos calcular várias métricas de desempenho que são utilizadas na avaliação da performance da classificação sendo estas a fração de verdadeiro positivo (sensibilidade), fração de falso positivo, precisão (especificidade) e acurácia. As métricas são dadas a seguir

Tabela 3 – Tabela representando a matriz de confusão

	Classe Verdadeira	
Classe predita	p	n
p^*	VP	FP
n^*	FN	VN
Total	$\mathbf{P}$	$\mathbf{N}$

Fonte: Gerada pelo autor.

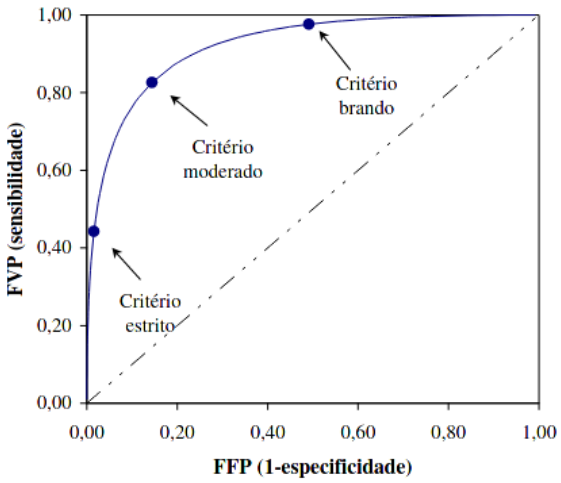
$$
\begin{aligned}
FVP &= \frac{VP}{VP+FN}, \\
Especificidade &= \frac{VN}{VN+FP}, \\
FFP &= 1 - Especificidade \\
&= \frac{FP}{VN+FP}, \\
Acurácia &= \frac{VP+VN}{\mathbf{P}+\mathbf{N}}.
\end{aligned}
$$

4.1.2 Curva ROC e medida de AUC

Por definição, uma curva ROC é uma representação gráfica das medidas de sensibilidade (FVP) e FFP (1-especificidade). Desta forma, a curva ROC representa de forma empírica a capacidade do sistema de diagnóstico de descriminar (classificar) entre as duas classes do conjunto $\{p, n\}$. Em termos de teste de hipótese temos duas distribuições para negativos e positivos onde podemos denotar como hipótese nula que a média da população é μ_0 representando a média da distribuição de negativos e a hipótese alternativa como sendo a média μ_1 como média da distribuição de positivos. Desta forma, uma curva ROC é conceitualmente equivalente a uma curva que mostra a relação entre a potência de teste e a probabilidade de cometer um erro de tipo I com a variação do "valor críitico" do teste estatístico (BRAGA, 2001).

Pode-se definir critérios correspondendo a pontos na curva ROC como por exemplo, designa-se o paciente como doente quando a evidência de doença é alta. Este critério conduz a uma taxa pequena de falsos positivos e também a uma pequena taxa de verdadeiros positivos. Pelo gráfico podemos ver o impacto de definição do critério, quanto mais estrito, maior as taxas de ambas FP e VP.

Figura 8 – Curva ROC com variação do critério de decisão.



Fonte: (BRAGA, 2001)

Para avaliação do desempenho de diferentes diagnósticos (modelos de classificação), temos que o que se aproxima mais do canto superior esquerdo é o que apresenta o maior poder de classificação (discriminação).

A medida de área abaixo da curva ROC conhecida como AUC representa o grau de medida de separabilidade das distribuições de probabilidade associadas a curva ROC, ou seja, o AUC nos diz o quanto o modelo é capaz de distinguir entre classes. Em avaliação de diagnósticos para doenças, por exemplo, quanto maior a AUC, melhor será a capacidade do modelo em distinguir entre pacientes com a doença e sem doença.

Um modelo excelente tem AUC próximo de 1, o que significa que tem uma boa medida de separabilidade. Um modelo pobre tem um AUC próximo a 0, o que significa que ele tem a pior medida de separabilidade. Na verdade, significa que está retribuindo o resultado. Ele está prevendo zeros como uns e uns como zeros. E quando AUC é $0,5$, significa que o modelo não tem capacidade de separação de classes.

4.2 CURVAS DE DECISÃO

Os modelos de previsão são geralmente avaliados aplicando-os a um conjunto de dados e comparando as previsões do modelo com o resultado real do paciente. Os resultados são normalmente expressos em termos de AUC como visto na sessão anterior.

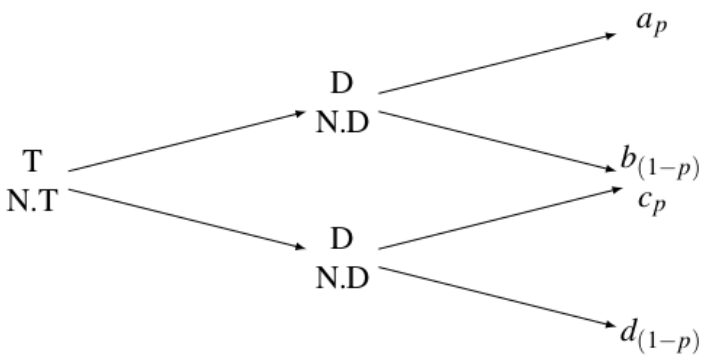
O AUC só nos informa acerca da precisão preditiva do modelo, e portanto não nos diz nada sobre as consequência de usar o modelo ou qual dos dois ou mais modelos é preferível. Para exemplificar uma consequência clara, considere o caso em que um resultado falso-negativo é muito mais prejudicial do que um resultado falso-positivo. Um modelo que tivesse uma especificidade muito maior, mas uma sensibilidade ligeiramente menor do que outro, teria uma AUC mais alta, mas seria uma escolha pior para uso clínico (HUNINK et al., 2014).

Os métodos com curvas de decisão nos permitem incorporar as consequência de um determinado diagnóstico, em teoria, podem nos dizer qual dentre os modelos alternativos deve ser usado (VICKERS; ELKIN, 2006). Em uma análise de decisão de modelos alternativos de diagnóstico, o modelo ótimo é aquele que maximiza o resultado de interesse. Foram propostas técnicas para simplificar as análises de decisão de testes diagnósticos usando uma relação risco-benefício para resumir os resultados de saúde associados às consequências dos testes (DJULBEGOVIC; DESOKY, 1996).

Um modelo de previsão não pode ser avaliado em uma análise de decisão sem que se tenha o resultado do modelo e o verdadeiro estado da doença ou resultado, estas não encontradas no conjunto de dados de validação. Essa é uma desvantagem da aplicabilidade desta metodologia.

Considere um problema em que um paciente deve decidir se submeter a um determinado tratamento, porém este não tem certeza se está doente ou não. Uma árvore de decisão simples pode ser construída em que p é a probabilidade de doença e a, b, c e d fornecem o valor associado a cada resultado em termos de anos de vida ajustados pela qualidade. O esquema a seguir mostra a árvore com T sendo tratamento, N.T o não tratamento, D a doença e N.D a não doença, sendo os valores a, b, c e d indexados pelas suas respectivas probabilidades:

Figura 9 – Árvore binomial.



Fonte: O autor (2022).

Considere que existe um modelo de previsão disponível. Isso fornece uma probabilidade de que o paciente tenha a doença: se a probabilidade da doença for próxima de 1, o paciente decidirá pelo tratamento; se a probabilidade for próxima de 0, é provável que não opte pelo tratamento. Com alguma probabilidade entre 0 e 1, o paciente não terá certeza se adere ao tratamento. Este limite de probabilidade, p_t, é quando o benefício esperado do tratamento é igual ao benefício esperado de evitar o tratamento. Resolvendo a árvore de decisão:

$$p_t a + (1 - p_t)b = p_t c + (1 - p_t)d,$$

que após uma simples álgebra, tem-se

$$\frac{a - c}{d - b} = \frac{1 - p_t}{p_t}. \tag{4.1}$$

Agora $(d - b)$ é a consequência de ser tratado desnecessariamente. Se o tratamento for orientado por um modelo de previsão, esse é o dano associado a um resultado falso-positivo (em comparação com um resultado verdadeiro-negativo). Comparativamente, $(a - c)$ é a consequência de evitar o tratamento quando ele teria sido benéfico, ou seja, o dano de um resultado falso-negativo (em comparação com um resultado verdadeiro-positivo) (VICKERS; ELKIN, 2006). Portanto, a equação (4.1) é a probabilidade limite de como é ponderado os danos relativos de resultados falso-positivos e falso-negativos pelo paciente.

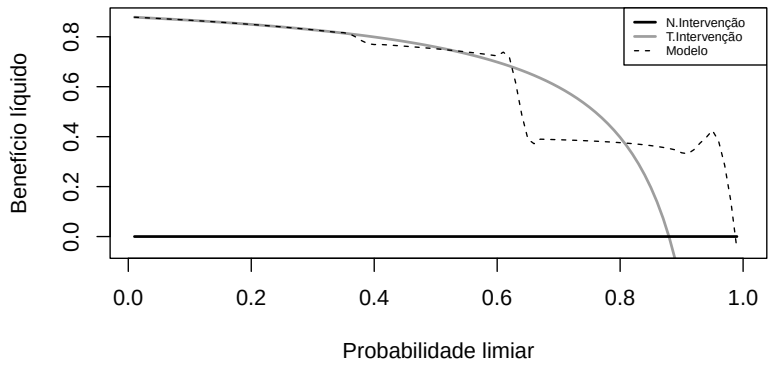
Os gráficos de curva de decisão portanto são construídos com o eixo x como sendo estas probabilidades limites (limiares) e no eixo y o benefício líquido relacionado com a frequência de falsos positivos e frequência de verdadeiros positivos. Para atribuir o valor do benefício líquido fixa-se o valor de $(a - c)$, o valor de um verdadeiro-positivo em 1, obtendo então o

valor de um falso-positivo como $p_t/(1-p_t)$. Desta forma, o benefício líquido pode ser expresso na forma

$$BL = \frac{FVP}{n} - \frac{FFP}{n}\left(\frac{p_t}{1-p_t}\right). \tag{4.2}$$

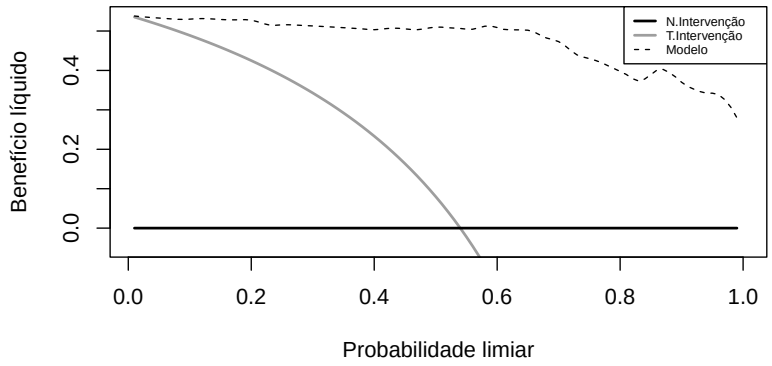
Em suma, subtrai-se a proporção de todos os pacientes que são falso-positivos da proporção que são verdadeiros positivos, ponderando pelo dano relativo de um resultado falso-positivo e um falso-negativo (VICKERS; ELKIN, 2006). Para comparação entre o modelo de classificação proposto para determinar uma intervenção usa-se a intervenção em todos os pacientes (geralmente isto é feito em doenças onde tem-se um tratamento fácil eficaz por medicamentos já estabelecido e comprovado), ou nenhuma intervenção (geralmente em doenças novas onde não se tem tratamentos eficazes comprovados). Desta forma, definindo uma probabilidade limiar pode-se ter a escolha do uso do modelo de classificação ou da abordagem de intervenção para todos de acordo com o benefício líquido, a escolha de um ou de outro é de acordo com o maior benefício líquido. Os gráficos a seguir representam duas situações em curvas de decisão

Figura 10 – Curva de decisão para um modelo genérico.



Fonte: O autor (2022).

Figura 11 – Curva de decisão para um modelo genérico.



Fonte: O autor (2022).

Entende-se que para valores mais a extrema esquerda do eixo x, ou seja, pequenas probabilidades de limiar, o profissional tem muita preocupação com a dano relativo não intervenção, e quando se define uma probabilidade de limiar mais próximo ao extremo direito do eixo x tem-se a situação onde o profissional pode considerar o dano muito alto para uma intervenção muito arriscada.

Temos assim um gráfico que pondera os danos relativos à intervenção ou não em uma doença e que pode auxiliar na tomada de decisão do profissional. Analisemos as curvas de decisão apresentadas anteriormente. Para a primeira curva na Figura 10 temos que o modelo não ajuda o médico a tomar qualquer decisão visto que o benefício líquido entre realizar intervenção em todos ou utilizar o modelo é o mesmo para uma ampla gama de probabilidades de limiar e o modelo é até inferior em benefício para uma determinada faixa entre $0,6$ a $0,8$. Já a curva do modelo apresentado na Figura 11 tem benefício líquido maior que a intervenção total para basicamente todas as probabilidades de limiar, este seria um modelo ideal na tomada de decisão do profissional de saúde para diferentes preferências preestabelecidas.

5 APLICAÇÃO

5.1 SIMULAÇÕES

Nesta seção realizaremos simulações com o objetivo de investigar como a imputação de dados pode afetar o viés e desvio padrão do erro da validação cruzada mediante a comparação das estimativas de desvio padrão da validação cruzada aninhada (VCA) (3.19) e viés (3.20) entre o erro de validação cruzada e validação cruzada aninhada em comparação com estas estimativas utilizando a análise de caso completo (quando exclui-se as observações faltantes). Simulamos dados onde $p > n$ e inserimos dados faltantes em escalas de 20% e 40% aproximadamente com tamanhos de amostras de 100, 500 e 1000. Toda a análise foi feita no software R e os modelos utilizados serão o LASSO, sendo ajustado pelo pacote glmnet, e o GaLASSO que é implementado no pacote miselect. A construção do algoritmo de VCA é feita com base no que é proposto em BATES; HASTIE; TIBSHIRANI (2021) com as modificações necessárias para trabalhar com dados imputados.

Os resultados são mostrados nas tabelas seguintes contendo o modelo, porcentagem de dados faltantes aproximada, a redução de tamanho de amostra para a análise de caso completo, para os campos com "-" significa que não houve redução do tamanho da amostra, n.r (n reduzido), EVCA, viés, desvio padrão (DP) e intervalo de confiança (IC) para EVCA ao nível $\alpha = 0,05$:

Tabela 4 – Resultados das avaliações de validação cruzada aninhada para os modelos Lasso e GaLasso com $n = 100$

Modelo	% falt. aprox.	n.r	EVCA	viés	DP	IC
LASSO	20%	76	$0,3323$	$-0,0390$	$0,4554$	$(0,1817; 0,4830)$
GaLASSO	20%	-	$4,8750$	$-2,0833$	$1,1964$	$(3,8270; 5,9243)$
LASSO	40%	76	$0,2733$	$-0,0271$	$0,4309$	$(0,1709; 0,3757)$
GaLASSO	40%	-	$6,7090$	$-5,3498$	$0,2650$	$(6,1894; 7,2285)$
LASSO	-	-	$0,2093$	$0,0256$	$0,4240$	$(0,1505; 0,2681)$

Fonte: O autor (2022).

Tabela 5 – Resultados das avaliações de validação cruzada aninhada para os modelos Lasso e GaLasso com $n = 500$

Modelo	% falt. aprox.	n.r	EVCA	viés	DP	IC
LASSO	20%	404	$0,2428$	$0,0297$	$0,4452$	$(0,1993;\ 0,2862)$
GaLASSO	20%	-	$10,2269$	$-8,7480$	$0,4411$	$(9,6568;10,7970)$
LASSO	40%	306	$0,3491$	$0,0377$	$0,4870$	$(0,2945;0,4037)$
GaLASSO	40%	-	$10,3990$	$-8,9783$	$0,2313$	$(10,2867;10,5112)$
LASSO	-	-	$0,1559$	$0,0060$	$0,3684$	$(0,1331;\ 0,1788)$

Fonte: O autor (2022).

Tabela 6 – Resultados das avaliações de validação cruzada aninhada para os modelos Lasso e GaLasso com $n = 1000$

Modelo	% falt. aprox.	n.r	EVCA	viés	DP	IC
LASSO	20%	820	$0,2989$	$0,0016$	$0,4585$	$(0,2667;0,3311)$
GaLASSO	20%	-	$13,9574$	$-12,6320$	$0,1078$	$(13,7460;14,1688)$
LASSO	40%	593	$0,3220$	$-0,0188$	$0,4672$	$(0,3032;0,3785)$
GaLASSO	40%	-	$14,2592$	$-12,9521$	$0,2826$	$(13,7052;14,8131)$
LASSO	-	-	$0,1198$	$0,0026$	$0,3278$	$(0,1054;\ 0,1342)$

Fonte: O autor (2022).

Temos alguns resultados interessantes para discussão. Em tamanhos de amostra pequenos, o impacto da retirada dos valores faltantes não foi tão alto o que pode ser pelo fato da amostra ser pequena o que faz com que em proporção não se tenha uma perca grande, o modelo GaLasso pode ser um pouco mais afetado pois a imputação PMMBB geralmente não é eficiente para bancos de dados pequenos, além disso, a diferença entre o erro de validação cruzada (EVC) e o EVCA já se mostra grande no GaLasso, este sendo apontado pelo viés alto em relação ao Lasso que pode ser visto na Tabela 4.

No cenário de 500 observações continuamos não tendo impacto evidente nos resultados de Lasso com exclusão dos dados faltantes e diminuição tamanho da amostra pelo menos para os resultados relacionados a erro de validação, viés e variância. Vemos que o viés é ainda maior entre o EVC e o EVCA no GaLasso, uma possível explicação pode estar no algoritmo do modelo GaLasso para matrizes com altas dimensões. Vemos que o desvio padrão do EVCA do Lasso e GaLasso é equiparada quando tem-se 20% de dados faltantes mas para 40% de dados faltantes esta é menor para o GaLasso. Os valores do Lasso são próximos tanto para o cenário com dados faltantes quanto para os dados completos sendo um pouco piores quando há dados faltantes mas a diferença não é tão alta.

Por fim, quando temos 1000 observações vemos que a estimativa de desvio padrão no GaLasso permanece menor, o viés entre EVC e EVCA se torna ainda maior, novamente os valores do Lasso com dados faltantes ou com a amostra completa continuam similares. Os erros de validação são maiores para o modelo GaLasso possivelmente pelas imputações de dados envolvidas pela introdução de certo erro pela tentativa de inserir dados que em sua essência são fictícios mesmo que tentem capturar um valor real subjacente, ou seja, conceitualmente estamos introduzindo certa incerteza quando se imputa os dados o que poderia aumentar a variância do erro, porém, quando estamos com tamanhos de amostra consideráveis podemos ver que a abordagem de imputação múltipla tem uma estimativa de desvio padrão pequena e o real impacto está no viés entre o EVC em relação ao EVCA, e neste caso, um viés tão alto impacta na qualidade desta estimativa de desvio padrão. Ainda sim, o GaLasso apresentou um melhor desempenho de classificação em comparação com o Lasso com a retirada de dados.

Vejamos agora como se comporta os modelos quanto ao seu poder de classificação. A tabela a seguir contém os valores de acurácia, sensibilidade, especificidade e AUC para cada modelo nos respectivos tamanhos de amostra

Tabela 7 – Resultados das medidas de precisão para Lasso com n reduzido, GaLasso e Lasso com n completo

Modelo	n	% falt.aprox	Sensibilidade	Especificidade	Acurácia	AUC
LASSO	76	20%	0,8246	0,3953	0,6400	0,6100
GaLASSO	100	20%	0,8421	0,4884	0,6900	0,6652
LASSO	76	40%	0,8070	0,4651	0,6600	0,6361
GaLASSO	100	40%	0,8421	0,4884	0,6900	0,6652
LASSO	100	-	0,9123	0,8605	0,8900	0,8864
LASSO	404	20%	0,7148	0,8373	0,7660	0,7760
GaLASSO	500	20%	0,7595	0,8469	0,7960	0,8032
LASSO	306	40%	0,6014	0,7751	0,6740	0,6882
GaLASSO	500	40%	0,7952	0,7271	0,7331	0,7312
LASSO	500	-	0,9553	0,9522	0,9540	0,9537
LASSO	820	20%	0,8426	0,7453	0,7960	0,7940
GaLASSO	1000	20%	0,8772	0,7808	0,8310	0,8290
LASSO	593	40%	0,8119	0,5720	0,6970	0,6920
GaLASSO	1000	40%	0,8599	0,7495	0,8070	0,8047
LASSO	1000	-	0,9827	0,9791	0,9810	0,9809

Fonte: O autor (2022).

Podemos notar pela tabela que em todos os valores de sensibilidade, especificidade, acurácia e AUC os GaLasso apresenta valores maiores que os do Lasso com amostra reduzida,

com diferenças de 3% a 4%. Além disso, os valores do GaLasso são mais próximos dos valores de Lasso com a amostra completa. Desta forma, o GaLasso apresenta um bom desempenho comparado ao Lasso mesmo com a introdução de viés com a imputação e esta parece ser uma alternativa melhor a retirada das observações faltantes. Na próxima seção aplicamos a abordagem em um banco de dados real onde temos $p < n$ porém com multicolinearidade e veremos na prática a diferença que faz a exclusão ou não das observações faltantes, em especial ao tamanho da amostra que torna-se bem reduzido.

Não iremos analisar curvas de decisão nas simulações pois estas precisam de uma definição precisa de benefício entre uma decisão de intervenção ou não em um contexto clínico.

5.2 APLICAÇÃO PARA UM BANCO DE DADOS REAL

No contexto clínico é comum a presença de dados faltantes devido a inúmeros fatores como perda de prontuários, exames incompletos por desistência do paciente, problemas de medição, entre outros. Desta forma, faz sentido a imputação de dados para estes bancos que notavelmente teriam esta observação subjacente ao paciente.

Em dezembro de 2019 tivemos primeiro caso conhecido da COVID-19 (nome dado à doença causada pelo vírus SARS-CoV-2) em Wuhan, na China. Em 20 de janeiro de 2020, a Organização Mundial da Saúde (OMS) classificou o surto como Emergência de Saúde Pública de Âmbito Internacional e, em 11 de março de 2020, como pandemia. Desde então pesquisadores de todo o mundo tem estudado a fundo a doença para poder buscar a melhor forma de tratar os pacientes acometidos da doença. A fase de intubação, por exemplo, é uma fase complicada e a decisão de se intubar o paciente ou não (quanto ao tempo de espera de iniciar a intubação e duração da intubação) deve ser feita de forma a maximizar a chance de recuperação do paciente (ALENCAR et al., 2021).

Uma forma de se entender mais sobre a doença é estudando e determinando as possíveis comorbidades ou variáveis de interesse clínico geral dos pacientes que podem influenciar no agravamento da doença. Isto pode ser feito através de um modelo de regressão onde a variável resposta de interesse, por exemplo, pode ser um desfecho de morte pela COVID-19, no nosso caso esta variável é a intubação, e as variáveis preditoras podem ser qualquer variável de interesse clínico sugeridas pelos pesquisadores, médicos e profissionais de saúde envolvidos.

Neste trabalho, os dados a serem trabalhados são de pacientes com COVID-19 com informações a respeito de comorbidades dos pacientes, exames, tratamento por intubação,

todos estes fornecidos pelo Grupo de Emergências do Hospital das Clinicas da USP que tem como objetivo a busca por biomarcadores de gravidade da doença que possam ser úteis para identificar pacientes mais graves que necessitem cuidados intensivos e se necessário a intubação (NETO et al., 2021).

5.2.1 Modelo de regularização nos dados multi-imputados

A variável resposta de interesse nesta análise é a intubação, esta sendo binária com 1 para intubado e 0 caso contrário. A COVID-19 afeta principalmente as vias respiratórias. O SARS-CoV-2 ataca o epitélio pulmonar, causando um processo inflamatório local que gera uma dificuldade na troca gasosa, podendo levar à falência respiratória. Esse fato é representado pela diminuição da oxigenação do sangue, identificada pela diminuição da saturação da hemoglobina. Assim, o parâmetro SatO2 representa o quanto a troca gasosa no pulmão está comprometida, e é uma das covariáveis de maior interesse de presença no modelo. Se a SatO2 for muito baixa, tem-se grande chance de se optar pela intubação do paciente (ventilação mecânica). O banco de dados tanto a variável resposta de interesse intubação quanto as covariáveis que serão consideradas possuem dados faltantes que precisam ser imputados. Desta forma, o modelo utilizado para classificação será o modelo logístico regularizado agrupado heterogêneo e homogêneo com penalização de grupo GaLASSO, e penalização empilhada SaENET. Também avaliaremos o modelo LASSO sem imputação. São 45 covariáveis avaliadas no modelo, embora $p < n$, as covariáveis tem correlação considerável apresentando assim multicolinearidade sendo necessário a implementação de modelos com regularização.

Desta forma, separamos o conjunto de dados com tamanho de amostra de 3596 pacientes em treino e teste, sendo 80% para treino e 20% para testar o poder preditivo. A proporção de intubados e não intubados antes da separação do banco de dados são de $47,52\%$ e $42,82\%$ respectivamente, o que é um balanceamento razoável, os demais $9,66\%$ são dados faltantes. A proporção no conjunto de treino se manteve próxima com $47,15\%$ para intubados e $42,94\%$ para não intubados com $9,91\%$ de dados faltantes, assim como no conjunto de teste que foi de $48,88\%$ para intubados, $42,36\%$ para não intubados e $8,76\%$ de dados faltantes. Com bases nestes conjuntos faremos as análises propostas.

O primeiro passo é a realização da imputação, esta será feita pelo método da correspondência média preditiva por boostrap bayesiano (BBPMM) introduzida no Capítulo 2, que no software R está implementado no pacote BaBooN. Para ajuste dos modelos será utilizado as

funções cv.galasso e sv.saenet do pacote miselect. As estimativas dos parâmetros de cada um dos modelos LASSO, GaLASSO e SaENET para a intubação são dadas na tabela 5.2.1, por simplicidade mantemos os termos em inglês para uma melhor organização da tabela. Podemos notar que o LASSO tem mais coeficientes iguais a zero, mantendo apenas 9 covariáveis no modelo, o modelo SaENET permaneceu com 31 covariáveis com coeficientes estimados diferentes de zero, e o modelo GaLASSO permaneceu com 12 covariáveis com coeficientes estimados diferentes de zero. As covariáveis concordantes nos três modelos são as drogas vasoativas (vasoactive drugs), saturação do oxigênio (SatO2), terapia de oxigênio (oxygen therapy), pcr 72 horas, última medição diastólica (diastolic bp last), última medição de frequência respiratória (respiratory rate last) e tempo de permanência na UTI (Length of stay in the ICU).

Tabela 8 – Estimativas dos parâmetros para os modelos de regularização nos dados imputados, m=5.

Covariável	LASSO	GaLASSO	SaENET
age	-	-0,0074	-0,0175
sex	-	-	0,0283
diabetes	-	-	0,0810
Length of hospital stay (days)	-	-	-0,0122
Length of stay in the ICU	0,0046	0,0753	0,1138
time symptoms admission	-	-	0,0187
temperature admission	-	-	0,0415
heart rate admission	-	-	-0,0084
respiratory rate admission	-	-	0,0092
systolic bp admission	-0,0018	-	-
diastolic bp admission	-	-	-0,0061
Oxygen saturation (%) on admission	-	-	-
weight admission	-	-	0,0014
height admission	-	-	0,0032
bmi admission	-	-	-
temperature last	-	0,1136	0,2186
heart rate last	-	-	0,0036
respiratory rate last	0,0028	0,0070	0,0161
systolic bp last	-	-	-
diastolic bp last	-0,0117	-0,0094	-0,0155
covid status	-	-	0,1296
leucocitos 72h	-	-	-
neutrofilos 72h 0,0186	0,0568	0,0876	
linfocitos 72h	-	-	-
hemoglobina 72h	-	-	-
hematocrito 72h	-	-	0,0053
plaquetas 72h	-	-	-
pcr 72h	0,0003	0,0012	0,0020
tgo 72h	-	-	0,0001
tgp 72h	-	-	-
tp 72h	-	-	-
ttpa 72h	-	-	0,0069
inr 72h	-	-	-
cpk 72h	-	-	-
lactato 72h	-	-	-
ureia 72h	-	-	-
creatinina 72h	-	-0,0332	-0,0983
sodio 72h	-	-	0,0286
potassio 72h	-	-	0,0727
calcio ionico 72h	-	-	0,1524
magnesio 72h	-	0,6623	0,8171
troponina 72h	-	-	-0,0011
vasoactive drugs	2,2839	3,4833	3,6684
oxygen therapy	2,2597	2,5005	2,9334
SatO2	0,0688	0,3281	0,4700

Fonte: O autor (2022).

As medidas de sensibilidade, especificidade, acurácia e AUC são dadas na Tabela 9 e podem nos mostrar como os modelos se comportam com relação ao poder de classificação

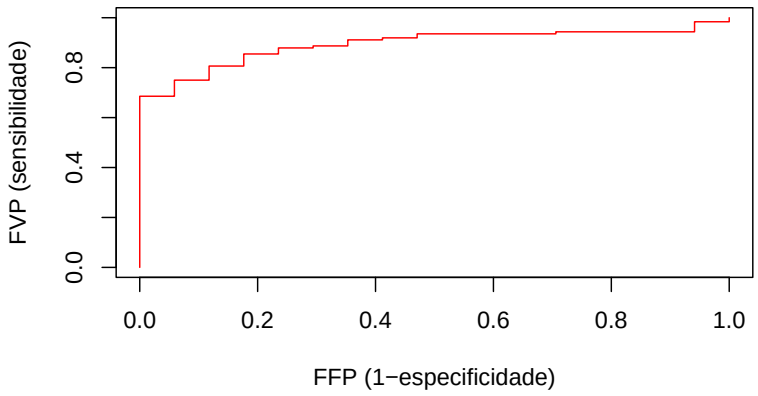
Tabela 9 – Medidas de desempenho de classificação dos modelos.

Medida	LASSO	GaLASSO	SaENET
Sensibilidade	0,9435	0,9111	0,9164
Especificidade	0,1176	0,9255	0,9226
Acurácia	0,8440	0,9181	0,9194
AUC	0,5306	0,9183	0,9195

Fonte: O autor (2022).

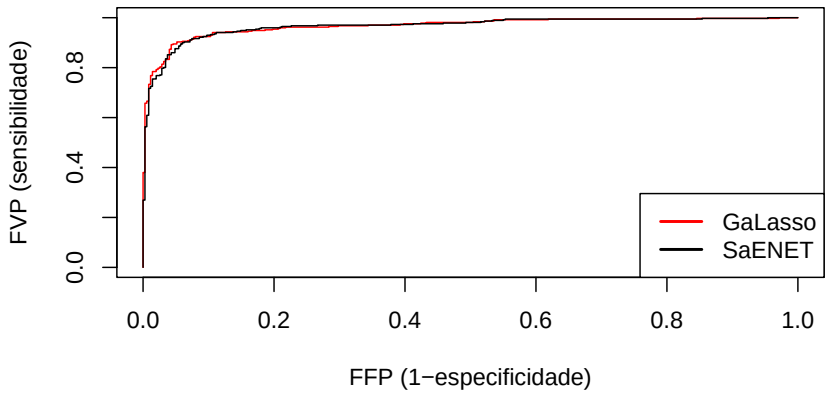
Podemos notar que a diferença entre a sensibilidade a especificidade no modelo de Lasso com a amostra reduzida é muito alta, o que pode desbalancear a acurácia, desta forma, a medida de AUC captura melhor o poder de classificação, desta forma, tivemos um poder de classificação baixo com relação aos modelo GaLASSO e SaENET. Olhando para os gráficos de curva ROC, note que o gráfico para o modelo de LASSO usa menos observações, o que muda a suavização da curva, podemos ver que os modelos de GaLASSO e SaENET tem área abaixo da curva maiores que a do Lasso

Figura 12 – Curva ROC do modelo LASSO.



Fonte: O autor (2022).

Figura 13 – Curvas ROC do modelo GaLASSO e SaENET.



Fonte: O autor (2022).

Vejamos agora as medidas de erro de validação cruzada aninhada com os respectivos valores de viés, desvio padrão e intervalos de confiança para os modelos, estes são dados na tabela a seguir

Tabela 10 – Resultados dos EVCA para os modelos LASSO, GaLASSO e SaENET.

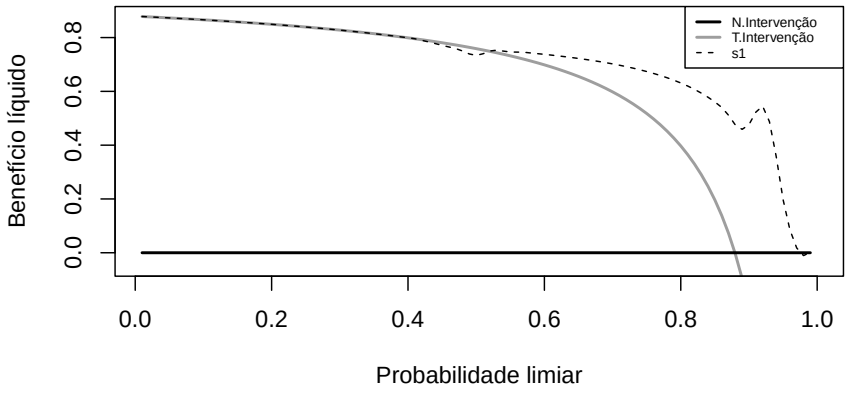
Modelo	% falt. aprox.	n.r	EVCA	viés	DP	IC
LASSO	-	716	$0,1278$	$-0,0110$	$0,3212$	$(0,1013;0,1543)$
GaLASSO	78%	-	$0,8533$	$-0,2917$	$0,1793$	$(0,6961;1,0105)$
SaENET	78%	-	$0,7933$	$-0,2813$	$0,1873$	$(0,6751;1,0127)$

Fonte: O autor (2022).

Podemos ver que o modelo de LASSO teve o menor EVCA e o menor viés com relação ao EVC, porém teve maior variância em relação ao GaLASSO e o SaENET. Novamente, temos que entender que o nosso banco de dados foi reduzido em 78% com a exclusão dos dados faltantes e o espectro de comparação pode não ser tão claro. Ainda sim, os modelos com imputação são bem promissores visto que mesmo com a imputação de valores fictícios, estes parecem capturar bem os valores faltantes subjacentes ao banco de dados. Neste caso, de fato é um grande ganho de tamanho amostral utilizar a metodologia de imputação sem um aumento muito grande do erro e desvio padrão para a construção do intervalo de confiança do erro como vistos no Capítulo 3.

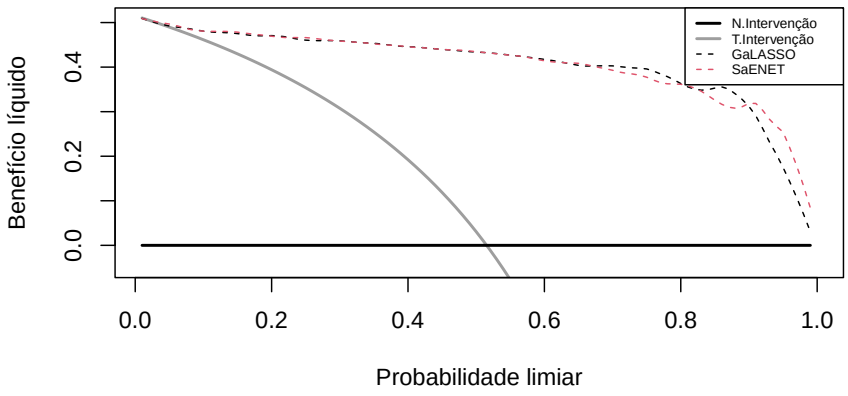
A medida de AUC na Tabela 9 deixou claro o impacto da retirada dos dados faltantes no poder de classificação. As curvas de decisão com relação aos modelos podem nos mostrar o impacto disto na decisão clínica da intubação ou não intubação de pacientes acometidos pela COVID-19. Os gráficos de decisão são dados a seguir para o modelo de Lasso (este com relação ao conjunto de teste reduzido) e os modelos de GaLASSO e SaENET

Figura 14 – Curva de decisão do modelo Lasso.



Fonte: O autor (2022).

Figura 15 – Curvas de decisão dos modelos GaLASSO e SaENET.



Fonte: O autor (2022).

Podemos ver que o modelo LASSO (s1) se iguala a intervenção total para uma grande

faixa de probabilidades de limiar, desta forma, o benefício líquido da intervenção total em relação ao uso do modelo para uma probabilidade de limiar por volta de $0,5$ é o mesmo, desta forma, a probabilidade limiar (preferência do médico) fica em faixas nas quais é difícil de decidir pela intubação. Neste caso, a intervenção total ou o uso do modelo para classificar quem deve receber intervenção não tem diferença, desta forma, não tem porque o médico ou profissional de saúde utilizar deste modelo como parâmetro de decisão.

Já para os modelos de GaLASSO e SaENET temos um benefício líquido bastante superior para a grande maioria das probabilidades limiares, sendo que este benefício só se iguala em probabilidades de limiar (preferências) muito baixas, estas geralmente são escolhidas por profissionais muito conservadores, ou que estejam extremamente preocupados com certas comorbidades do paciente em questão e os possíveis efeitos da não intubação. Na maioria das decisões por parte do profissional de saúde, mesmo com probabilidades de limiar baixas (alta preocupação com a não intervenção) os modelos se saem melhor para decisão em relação a intervenção total sendo bem claro a tomada de decisão do médico com base nos dois modelos de classificação.

As curvas de decisão juntamente com a medida de AUC evidenciam bem o impacto da metodologia utilizada e respectivos modelos para geração de gráficos que tenham algum impacto positivo no auxilio da tomada de decisão do profissional de saúde e do paciente.

6 CONCLUSÃO

As simulações do Capítulo 5 mostraram aspectos problemáticos a se considerar com relação ao modelo GaLASSO. Na situação onde temos o número de covariáveis maior que o tamanho da amostra $p > n$, os resultados do viés entre o erro da validação cruzada aninhada e o erro de validação cruzada foram bastante altos, e com o aumento do tamanho da amostra (necessariamente aumento da quantidade de covariáveis) este viés continuou crescendo, em contra partida o desvio padrão diminuiu. Este viés impacta diretamente na qualidade da estimativa de desvio padrão com a validação cruzada aninhada pois precisamos de um baixo viés entre um erro e outo para uma boa estimativa de desvio padrão e construção de intervalos de confiança adequados. O modelo de LASSO para nossas simulações apresentou certa consistência mesmo com exclusão de dados sem muita diferença em relação as estimativas de erro de validação cruzada aninhada, viés e desvio padrão, mesmo estes sendo menores para a amostra completa como visto nas Tabelas 4, 5 e 7. Vimos que os valores de viés do modelo GaLASSO sofreram bastante com a alta dimensionalidade porém não ficou claro como ele foi afetado, se na imputação ou no algoritmo de ajuste do modelo proposto recentemente. Estes questionamentos podem ser uma possível direção em investigações futuras para, além da identificação do problema, propor possíveis soluções nestes cenários já que em baixa dimensionalidade não tivemos viés alto como vimos na abordagem com dados reais. Outro aspecto interessante é que mesmo com estes problemas o GaLASSO mostrou medidas de poder de classificação melhores que o Lasso com a amostra reduzida, estes sendo mais próximos do Lasso com a amostra completa, ou seja, mesmo com alto viés no erro de validação a classificação do GaLASSO foi melhor em comparação com a abordagem de retirada de dados faltantes.

Na aplicação prática em dados reais o cenário se diferencia por não termos $p > n$, mas multicolinearidade entre as covariáveis. A diferença entre o uso das metodologias ficaram mais evidentes no exemplo prático com os dados de COVID do HC-USP, tivemos diferenças significativas de poder de classificação dos modelos de LASSO com redução da amostra por meio da exclusão dos dados e da imputação dos dados faltantes e ajuste do modelo GaLASSO. O uso de um ou de outro teve impacto direto na decisão do profissional de saúde por meio da curva de decisão. Utilizando apenas o Lasso com a amostra reduzida a curva de decisão ficou inutilizável e não teve serventia para a tomada de decisão pois entre usar o modelo para determinar a intubação do paciente ou decidir intubar todos não tinha diferença de benefício

líquido para uma gama muito grande de probabilidades de limiar não tendo nenhum benefício prático. Utilizar-se da imputação múltipla e de modelos de penalização de grupo teve benefício prático no final das contas nos gráficos de curva de decisão, sendo preferível utiliza-los para praticamente toda a faixa de probabilidade limiar (preferência) do profissional, sendo de grande auxilio na decisão final do profissional.

Além disso, os valores de erro de validação cruzada aninhada, viés e desvio padrão ficaram pequenos e próximos entre os modelos como pode ser visto na Tabela 10, diferente do que foi obtido nas simulações. Novamente, é necessário investigar mais afundo como a alta dimensionalidade impacta na imputação ou no ajuste do modelo GaLASSO visto que tivemos ótimos resultados quando $p < n$.

Por fim, existem aspectos a considerar quando temos dados faltantes, as técnicas de imputação bayesianas tem um grande apelo computacional dependendo do tamanho da amostra, número de covariáveis e proporção de dados faltantes sendo um tempo considerável de execução em computadores um pouco mais modestos. Quando há bons recursos computacionais, a imputação se mostra viável e até necessária em situações como os dados do HC-USP onde a exclusão dos dados acarretava em perda de 78% do tamanho da amostra, e o impacto disso vimos no poder de classificação e na construção das curvas de decisão. É de suma importância definir a técnica que seja adequada para o cenário com que está se trabalhando, ponderando como a perda de informação da escolha de exclusão dos dados pode afetar as inferências e resultados. Os ganhos de precisão de classificação para definição de prognóstico com a imputação foram consideráveis e de fato os modelos com as imputações foram eficientes e úteis na tomada de decisão final do profissional que é o objetivo do uso de tais modelos.

REFERÊNCIAS

ALENCAR, J. C. G. de; STERNLICHT, J. M.; VEIGA, A. D. M.; MARCHINI, J. F. M.; FERREIRA, J. C.; CARVALHO, C. R. R.; MARCILIO, I.; SILVA, K. R.; JUNIOR, V. C.; FELIX, M. C. et al. Timing to intubation covid-19 patients: Can we put it off until tomorrow? 2021.

ALLISON, P. Imputation by predictive mean matching: Promise & peril. *Statistical Horizons*, 2015.

BATES, S.; HASTIE, T.; TIBSHIRANI, R. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 2021.

BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, v. 5, n. Sep, p. 1089–1105, 2004.

BRAGA, A. Curvas roc: aspectos funcionais e aplicações. 2001.

BUUREN, S. V. *Flexible imputation of missing data*. [S.l.]: CRC press, 2018.

BUUREN, S. V.; BRAND, J. P.; GROOTHUIS-OUDSHOORN, C. G.; RUBIN, D. B. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, Taylor & Francis, v. 76, n. 12, p. 1049–1064, 2006.

CHEN, Q.; WANG, S. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in medicine*, Wiley Online Library, v. 32, n. 21, p. 3646–3659, 2013.

DJULBEGOVIC, B.; DESOKY, A. H. Equation and nomogram for calculation of testing and treatment thresholds. *Medical Decision Making*, Sage Publications Sage CA: Thousand Oaks, CA, v. 16, n. 2, p. 198–199, 1996.

DU, J.; BOSS, J.; HAN, P.; BEESLEY, L. J.; GOUTMAN, S. A.; BATTERMAN, S.; FELDMAN, E. L.; MUKHERJEE, B. Variable selection with multiply-imputed datasets: choosing between stacked and grouped methods. *arXiv preprint arXiv:2003.07398*, 2020.

EFRON, B. Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, SIAM, v. 21, n. 4, p. 460–480, 1979.

EFRON, B. How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, Taylor & Francis, v. 81, n. 394, p. 461–470, 1986.

EFRON, B.; GONG, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, Taylor & Francis, v. 37, n. 1, p. 36–48, 1983.

EFRON, B.; TIBSHIRANI, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, Taylor & Francis, v. 92, n. 438, p. 548–560, 1997.

EGAN, J. P.; EGAN, J. P. *Signal detection theory and ROC-analysis*. [S.l.]: Academic press, 1975.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical learning with sparsity: the lasso and generalizations*. [S.l.]: Chapman and Hall/CRC, 2019.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

HUNINK, M. M.; WEINSTEIN, M. C.; WITTENBERG, E.; DRUMMOND, M. F.; PLISKIN, J. S.; WONG, J. B.; GLASZIOU, P. P. *Decision making in health and medicine: integrating evidence and values*. [S.l.]: Cambridge university press, 2014.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020.

KOLLER-MEINFELDER, F. Analysis of incomplete survey data-multiple imputation via bayesian bootstrap predictive mean matching. opus, 2009.

LITTLE, R. J. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 6, n. 3, p. 287–296, 1988.

LITTLE, R. J.; RUBIN, D. B. Bayes and multiple imputation. *Statistical analysis with missing data*, Wiley Online Library, p. 200–220, 2002.

LITTLE, R. J.; RUBIN, D. B. *Statistical analysis with missing data*. [S.l.]: John Wiley & Sons, 2019. v. 793.

MIOT, H. A. Valores anômalos e dados faltantes em estudos clínicos e experimentais. *Jornal Vascular Brasileiro*, SciELO Brasil, v. 18, 2019.

NETO, R. A. B.; MARCHINI, J. F.; MARINO, L. O.; ALENCAR, J. C.; NETO, F. L.; RIBEIRO, S.; SALVETTI, F. V.; RAHHAL, H.; GOMEZ, L. M. G.; BUENO, C. G. et al. Mortality and other outcomes of patients with coronavirus disease pneumonia admitted to the emergency department: A prospective observational brazilian study. *PloS one*, Public Library of Science San Francisco, CA USA, v. 16, n. 1, p. e0244532, 2021.

NUNES, L. N. Métodos de imputação de dados aplicados na área da saúde. 2007.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cadernos de Saúde Pública*, SciELO Brasil, v. 25, p. 268–278, 2009.

RUBIN, D. B. The bayesian bootstrap. *The annals of statistics*, JSTOR, p. 130–134, 1981.

RUBIN, D. B. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 4, n. 1, p. 87–94, 1986.

RUBIN, D. B. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, JSTOR, v. 82, n. 398, p. 543–546, 1987.

RUBIN, D. B. An overview of multiple imputation. In: CITESEER. *Proceedings of the survey research methods section of the American statistical association*. [S.l.], 1988. p. 79–84.

RUBIN, D. B.; SCHENKER, N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American statistical Association*, Taylor & Francis, v. 81, n. 394, p. 366–374, 1986.

SCHAFER, J. L. *Analysis of incomplete multivariate data*. [S.l.]: CRC press, 1997.

SWETS, J. A. Measuring the accuracy of diagnostic systems. *Science*, American Association for the Advancement of Science, v. 240, n. 4857, p. 1285–1293, 1988.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

VICKERS, A. J.; ELKIN, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, Sage Publications Sage CA: Thousand Oaks, CA, v. 26, n. 6, p. 565–574, 2006.