**UNIVERSIDADE FEDERAL DE JUIZ DE FORA**

**INSTITUTO DE CIÊNCIAS EXATAS**

**DEPARTAMENTO DE ESTATÍSTICA**

**ISABELA LOPES PENNA**

# SELEÇÃO DE MODELOS DE REGRESSÃO LINEAR EM BASES DE ALTA DIMENSÃO

Juiz de Fora

2021

**ISABELA LOPES PENNA**

# SELEÇÃO DE MODELOS DE REGRESSÃO LINEAR EM BASES DE ALTA DIMENSÃO

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Clécio da Silva Ferreira

Juiz de Fora

2021

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Penna, Isabela Lopes.
Seleção De Modelos De Regressão Linear Em Bases De Alta Dimensão / Isabela Lopes Penna. -- 2021.
45 f.

Orientador: Clécio da Silva Ferreira
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2021.

1. Seleção de Variáveis Regressoras. 2. Regressão Ridge. 3. Regressão LASSO. 4. Regressão de Componentes Principais. 5. Mínimos Quadrados Parciais. I. Ferreira, Clécio da Silva, orient. II. Título.

**ISABELA LOPES PENNA**

# SELEÇÃO DE MODELOS DE REGRESSÃO LINEAR EM BASES DE ALTA DIMENSÃO

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em 17/03/2021

BANCA EXAMINADORA

_____

Prof. Dr. Clécio da Silva Ferreira - Orientador
Universidade Federal de Juiz de Fora

_____

Professora Dra. Camila Borelli Zeller
Universidade Federal de Juiz de Fora

_____

Prof. Dr. Tiago Maia Magalhães
Universidade Federal de Juiz de Fora

Dedico este trabalho à minha querida avó Alzira (in memoriam), que sempre me apoiou e me ensinou a nunca desistir de meus sonhos.

## AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, que sempre me iluminou e me manteve forte durante esse ciclo. Sem Deus nada disso seria possível, nos momentos mais difíceis Ele me sustentou e me deu forças para continuar.

Agradeço aos meus pais, Laumir e Gisele, meu irmão Lauro, meu namorado Phelipe e minha avó Alzira (in memoriam), que sempre me apoiaram, me estimularam, me incentivaram e acreditaram em meu potencial. Obrigada por me ensinarem a persistir mesmo nas situações mais adversas, por me mostrarem que nem sempre a vida será gentil, mas com fé e coragem podemos recomeçar e superar os desafios.

Agradeço também aos meus familiares e amigos que de alguma forma contribuíram nessa caminhada, que me ajudaram, me deram forças e torceram por mim. Aos meus amigos que tive a oportunidade de conhecer durante a faculdade, o meu muito obrigado, vocês tornaram esse processo mais leve.

Agradeço aos meus professores, que compartilharam seus conhecimentos para que me tornasse uma profissional de excelência. Em especial, agradeço ao meu orientador Clécio Ferreira, por todo suporte, incentivo e disponibilidade, não apenas neste Trabalho de Conclusão de Curso, mas durante esses longos anos de faculdade.

## RESUMO

Em diversos casos é preciso lidar com bases de dados com um grande número de variáveis, bases essas denominadas Big Data. Em algumas situações se faz necessário reduzir o número dessas variáveis, seja para se ter um modelo apenas com as variáveis significativas ou para sanar problemas como multicolinearidade. Com isso surgiram vários métodos para solucionar os casos acima. Esse trabalho se propôs a estudar os diferentes métodos de seleção em modelos de regressão linear múltipla, como métodos de seleção de subconjuntos de variáveis, métodos de regularização e métodos de redução de dimensionalidade. Procedimentos de validação cruzada foram utilizados nos processos de regularização para seleção dos modelos. Aplicações em dados reais foram realizadas utilizando o *software* R Core Team (2018).



Palavras-chave: Seleção de Variáveis Regressoras. Regressão Ridge. Regressão LASSO. Regressão de Componentes Principais. Mínimos Quadrados Parciais.

## ABSTRACT

In many cases, it is necessary to deal with databases with a large number of variables, bases called Big Data. In some situations it is necessary to reduce the number of these variables, either to have a model with only the significant variables or to solve problems such as multicollinearity. With this, several methods have emerged to solve the above cases. This work proposed to study the different selection methods in multiple linear regression models, such as methods of selecting subsets of variables, methods of regularization and methods of dimensionality reduction. Cross-validation procedures were used in the regularization processes for selecting the models. Real data applications were performed using the R Core Team software (2018).



Keywords: Selection of Regressor Variables. Ridge Regression. LASSO Regression. Principal Component Regression. Partial Least Squares.

**LISTA DE ILUSTRAÇÕES**

## LISTA DE TABELAS

**SUMÁRIO**

## 1 INTRODUÇÃO

Quando se tem o objetivo de explicar uma variável em função de outras, podemos utilizar os modelos de regressão linear (MRL). Tais modelos assumem que o relacionamento entre a variável resposta e as variáveis explicativas é representado da seguinte forma:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_\mathrm{P} \mathrm{X}_\mathrm{P} + \epsilon\,,$$

em que $\mathrm{Y}$ denota a variável resposta; $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{p}$ denotam as variáveis explicativas ou regressoras, $\beta = (\beta_0, \beta_1, \ldots, \beta_\mathrm{p})$ é o vetor dos coeficientes de regressão relativo às variáveis explicativas; e $\epsilon$ é o erro aleatório.

Nos últimos tempos os bancos de dados se tornaram bancos complexos, com muitas variáveis e\ou observações, devido o avanço tecnológico e a gama de informações disponíveis. Esses bancos de dados de alta dimensão, denominados Big Data, precisam de uma abordagem específica quando se tem o objetivo de fazer um modelo de regressão linear com suas variáveis.

Um dos problemas que podem surgir é na estimação dos coeficientes de regressão. Como esses bancos de dados possuem uma grande quantidade de variáveis, elas podem apresentar multicolinearidade e então o método de mínimos quadrados, que é o método usual para estimar os coeficientes, não se apresenta eficiente.

Outro problema que aparece é no momento de fazer seleção de variáveis. Considerando $p$ o número de variáveis explicativas, existem $2^\mathrm{p}$ modelos possíveis para analisar quais variáveis são importantes para o modelo final, ou no caso de inferência paramétrica, quais variáveis são significantes para o modelo. Como são muitas variáveis, teríamos um alto custo computacional e também dificuldade em analisar cada um dos modelos.

Visando resolver os problemas acima, determinados métodos foram propostos. Para seleção de variáveis, alguns métodos disponíveis são: *Forward Selection*, *Backward Elimination* e *Stepwise Regression.* Aplicações desses métodos podem ser encontradas em Charnet et al (1999), Gangi et al. (2019), entre outros.

Sobre métodos de regularização, dois principais são a regressão Ridge e LASSO (Least Absolute Shrinkage And Selection Operator ), sendo o primeiro mais utilizado em situações onde existe o problema de multicolinearidade entre as variáveis explicativas e sem o objetivo de selecionar variáveis. O segundo além de sanar a multicolinearidade, também atua como método de seleção de variáveis. Aplicações desses métodos podem ser vistas em Reynaldo (1997), Pereira (2017), Melkumova e Shatskikh (2017). Neste último trabalho foram exploradas propriedades do vinho tinto com o objetivo de analisar e compreender quais características físico-químicas podem afetar a qualidade do vinho utilizando os métodos Ridge e LASSO para ajustar modelos visando prever a qualidade do vinho.

Por fim foram estudados dois métodos de redução de dimensionalidade: PCR (Regressão De Componentes Principais) e PLS (Mínimos Quadrados Parciais). Aplicações desses métodos podem ser encontradas em Haaland (1988), Hastie, Tibshirani e Friedman (2009), Mateos-Aparicio (2011).

Esse trabalho tem como objetivo estudar cada um desses métodos, entendendo suas particularidades. Após o estudo de cada método, serão utilizadas bases de dados reais para a aplicação dos mesmos, em que toda a parte de aplicação será desenvolvida no software R Core Team (2018).

# 2 REGRESSÃO LINEAR MÚLTIPLA

O termo regressão, de acordo com Rodrigues (2012), foi proposto pela primeira vez por Sir Francis Galton[1] em 1889, quando o estudioso demonstrou que a altura dos filhos não tende a refletir a altura dos pais, mas tende sim a regredir para a média da população. Hoje, a regressão tem por objetivo explicar um fenômeno, denominado variável resposta, através de outras variáveis, chamadas de variáveis explicativas ou regressoras. Esse modelo apresenta uma parte determinística e uma parte aleatória, de forma que podemos modelar relações entre variáveis e predizer o valor da variável resposta em função de um conjunto de variáveis regressoras.

Sob essa perspectiva, neste capítulo vamos introduzir o modelo de regressão linear múltipla. Para tanto, será apresentado o modelo teórico e os seus pressupostos, bem como a estimação dos parâmetros do modelo pelo método dos mínimos quadrados. Além disso, vamos abordar alguns conceitos, como a escolha de variáveis regressoras para serem incluídas no modelo, a multicolinearidade e porque esses conceitos podem se tornar um problema em bases de alta dimensão.

## 2.1 O MODELO

De acordo com Charnet et al (1999, p.170-174) e Rodrigues (2012, p.23-25), a equação do modelo de regressão linear múltipla com $p$ variáveis regressoras é dada por:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP} + \epsilon_i, \qquad i = 1, \dots, n.$$

Nessa fórmula, $Y_i$ representa a variável resposta na observação $i$; $x_{i1}, x_{i2}, \dots, x_{ip}$ são os valores da i-ésima observação das $p$ variáveis regressoras, sendo que essas variáveis são fixas e conhecidas. Ademais, temos que $\beta_0, \beta_1, \dots, \beta_p$ são os parâmetros do modelo, sendo esses parâmetros lineares, constantes e desconhecidos e $\epsilon_i$ correspondem aos erros aleatórios.

Esse modelo pode ser escrito em notação matricial da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

[1] Ver: Galton, F. Natural Inheritance. London, 1889.

onde $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ e $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$.

Sendo que:

- $\boldsymbol{Y}$ é um vetor $n\,x\,1$, constituído pelas observações das variáveis respostas;
- $\boldsymbol{X}$ é uma matriz de dimensão $n\,x\,(p+1)$, onde a primeira coluna é composta por 1's para acompanhar $\beta_0$ e as demais colunas representam as observações de cada variável regressora;
- $\boldsymbol{\beta}$ é um vetor coluna $(p+1)\,x\,1$ cujos elementos são os parâmetros do modelo, chamados também de coeficientes de regressão;
- $\boldsymbol{\epsilon}$ é um vetor de dimensão $n\,x\,1$, constituído pelos erros aleatórios.

Os pressupostos usuais para esse modelo são:

1. $E[\epsilon_i] = 0, i = 1, \ldots, n.$
2. Não existe correlação entre os erros das observações.
3. $Var[\epsilon_i] = \sigma^2, i = 1, \ldots, n.$

## 2.2 O MÉTODO DOS MÍNIMOS QUADRADOS E A ESTIMAÇÃO DOS PARÂMETROS

Um dos métodos mais utilizados para estimação dos coeficientes de regressão é o método dos mínimos quadrados, conforme abordado por Charnet et al (1999, p. 175-181) esse método consiste em estimar os parâmetros do modelo tentando minimizar a soma dos quadrados das diferenças entre os valores estimados ($\hat{y}_i$) e os dados observados ($y_i$). Essa diferença é chamada de resíduos.

O método dos mínimos quadrados propõe, então, encontrar os valores dos $\widehat{\boldsymbol{\beta}}'s$ para os quais a soma dos quadrados dos resíduos (SQR) é mínima:

$$SQR(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$$
$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}.$$

A partir desse ponto, derivando $\boldsymbol{\beta}$ obtemos:

$$\frac{\partial SQR}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}.$$

E igualando a derivada a zero, obtemos $\widehat{\boldsymbol{\beta}}$ dado por:

$$\begin{aligned} &-2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0 \\ &\leftrightarrow (\boldsymbol{X}^T\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y} \\ &\leftrightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \end{aligned}$$

Portanto, o estimador de mínimos quadrados de $\boldsymbol{\beta}$ é $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Sendo que só é possível extrair esse resultado quando a inversa de $\boldsymbol{X}^T\boldsymbol{X}$ exista. Além disso, podemos também estimar variância de $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned} Var[\widehat{\boldsymbol{\beta}}] &= Var[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Var[\boldsymbol{Y}][(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned}$$

Também é importante considerar que os estimadores de mínimos quadrados $\widehat{\boldsymbol{\beta}}'\boldsymbol{s}$ possuem importantes propriedades, pois eles são não viciados. Ou seja, $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ e tem variância mínima entre todos os estimadores não viciados que são combinações lineares dos $Y_i$.

Um estimador não viciado de $\sigma^2$ é dado pela soma de quadrados dos resíduos (SQR), dividida por $(n - p - 1)$, sendo $(p + 1)$ o posto da matriz $\boldsymbol{X}$:

$$\hat{\sigma}^2 = \frac{\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{n - p - 1}.$$

## 2.3 O MÉTODO DE MÁXIMA VEROSSIMILHANÇA E A ESTIMAÇÃO DOS PARÂMETROS

Outro método que também pode ser utilizado para a estimação dos parâmetros do MRLM é o método de máxima verossimilhança, de acordo com Montgomery, Peck e Vining (2012, p. 83-84). Nesse método de estimação é preciso supor uma distribuição para os erros ($\boldsymbol{\epsilon}$). Para este trabalho foi suposto que os erros do modelo são normais e independentemente distribuídos. Fazendo isso, é possível mostrar que os estimadores de máxima verossimilhança para os parâmetros do modelo na regressão linear múltipla também são estimadores de mínimos quadrados.

O MRLM pode ser reescrito da seguinte forma:

$$Y_i = \mu_i + \epsilon_i, \qquad i = 1, \dots, n,$$

em que $\mu_i = \boldsymbol{x_i}^T\boldsymbol{\beta}$, sendo $\boldsymbol{x_i}^T$ = [1 $x_{i1}$ $x_{i2}$ ... $x_{ip}$] a *i*-ésima observação das variáveis regressoras.

Entretanto, agora os erros seguem uma distribuição Normal e são independentemente distribuídos, com variância constante $\sigma^2$, ou seja, $\epsilon_i \sim N(0, \sigma^2)$. Portanto, tem-se que $Y_i \sim N(\mu_i, \sigma^2)$ e a função de verossimilhança é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{i=1}^{n} f_{\boldsymbol{y}}(y_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \emptyset(y_i; \mu_i, \sigma^2), \qquad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix},$$

onde $\emptyset(\cdot\,; \mu; \sigma^2)$ é a função densidade de probabilidade da Normal no com média $\mu$ e variância $\sigma^2$.

E então a função de log-verossimilhança é apresentada por:

$$l(\boldsymbol{\theta}) = \ln\big(L(\boldsymbol{\theta}, \boldsymbol{y})\big) = ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2.$$

A função de log-verossimilhança é maximizada quando o termo seguinte é minimizado:

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}).$$

Portanto, o estimador de máxima verossimilhança de $\boldsymbol{\beta}$ considerando os erros como normais é equivalente ao estimador de mínimos quadrados $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ e o estimador de $\sigma^2$ é:

$$\hat{\sigma}^2 = \frac{\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{n}.$$

## 2.4 SELEÇÃO DE VARIÁVEIS

Em determinados casos é possível ter um grande número de variáveis regressoras, que podem ser candidatas para explicar um fenômeno, também chamada variável resposta. Então surge o problema da escolha de variáveis regressoras para compor o modelo, isto é, decidir se devem incluir todas as variáveis regressoras disponíveis ou incluir apenas um subconjunto dessas variáveis ao modelo.

Outrossim, não existe um método exclusivo para a seleção do melhor subconjunto de variáveis. De acordo com Charnet et al (1999, p. 264-271), serão abordados aqui alguns métodos para essa seleção. São estes: *Forward Selection*, *Backward Elimination* e *Stepwise Regression*. Dentro de cada método é usado um critério para a eliminação/adição ou não da variável, e neste trabalho abordaremos três deles: AIC, BIC e Estatística F.

### 2.4.1 *FORWARD SELECTION* (“PASSO A FRENTE”)

Esse procedimento inicia-se com a suposição de que não há regressores no modelo, a primeira variável a entrar no modelo é aquela de maior coeficiente de correlação amostral observado com a variável resposta. Começa então o processo de adição de uma variável por vez ao modelo de acordo com o critério escolhido. Se

em uma etapa não houver a inclusão de nenhuma variável, o processo é interrompido e as variáveis selecionadas até esta etapa definem o modelo final.

### 2.4.2 *BACKWARD ELIMINATION* (“PASSO ATRÁS”)

Esse procedimento caracteriza-se por incorporar, inicialmente, todas as variáveis regressoras ao modelo. Inicia-se então o processo de eliminação de uma variável por vez de acordo com o critério escolhido. Se em uma etapa nenhuma variável for eliminada, o processo é interrompido e as variáveis restantes definem o modelo final.

### 2.4.3 *STEPWISE REGRESSION* (“PASSO A PASSO”)

Este procedimento é uma combinação das duas categorias acima, podendo começar tanto com um modelo sem nenhuma variável regressora ou com um modelo com todas as variáveis regressoras. Essa regressão requer dois valores de corte, um para adicionar variáveis e um para removê-las.

### 2.4.4 CRITÉRIO DE INFORMAÇÃO DE AKAIKE-AIC

Akaike (1974) desenvolveu uma estimativa de informação baseada na Função de Log-Verossimilhança em seu ponto máximo, acrescida de uma penalidade associada ao número de parâmetro do modelo, conhecida como AIC. Uma medida que estima a qualidade de cada um dos modelos possíveis, sendo o melhor modelo àquele que tem o menor AIC. O Critério de Informação de Akaike é definido pela seguinte fórmula:

$$\mathrm{AIC} = -2\mathrm{l}\left(\widehat{\boldsymbol{\theta}}\right) + 2(\mathrm{p}+2),$$

no qual $p$ é o número de variáveis regressores no modelo e $\mathrm{l}\left(\widehat{\boldsymbol{\theta}}\right)$ é a função de log-verossimilhança do modelo.

### 2.4.5 CRITÉRIO DE INFORMAÇÃO BAYESIANO-BIC

O Critério de Informação Bayesiano (BIC), proposto por Schwarz (1978) é também uma medida de qualidade, sendo o melhor modelo àquele que tem o menor BIC, mas possui uma penalidade maior para o número de parâmetros do que o AIC. O Critério de Informação Bayesiano é definido como:

$$BIC = -2\mathrm{l}(\widehat{\boldsymbol{\theta}}) + (p+2)ln(n),$$

em que $p$ é o número de variáveis regressores no modelo, $n$ é o tamanho da amostra e $\mathrm{l}(\widehat{\boldsymbol{\theta}})$ é a função de log-verossimilhança do modelo.

### 2.4.6 ESTATÍSTICA F

A estatística F, usada por Charnet et al (1999), compara o modelo completo com o modelo reduzido, testando a contribuição de cada variável no modelo. Essa estatística é definida como

$$F = \frac{SQReg^{c} - SQReg^{r}}{\widehat{\sigma}^{2}},$$

em que $\mathrm{SQReg} = \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathrm{n}\bar{\mathbf{y}}^{2}$ é a soma de quadrados da regressão, sendo $\mathrm{SQReg}^{\mathrm{c}}$ a soma de quadrados da regressão do modelo completo e $\mathrm{SQReg}^{\mathrm{r}}$ a soma de quadrados da regressão do modelo reduzido, $\hat{\sigma}^{2}$ o estimador de variância do modelo completo.

Esse critério é utilizado conforme o procedimento escolhido. Portanto, para o método *Forward Selection* se começa com o modelo reduzido e acrescenta-se nesse modelo uma variável por vez, a cada variável adicionada tem-se um modelo completo e é possível calcular a estatística F de cada um deles e dentre essas é usada a maior ($\mathrm{F}_{\max}$) para comparar com o quantil especificado da distribuição F com 1 e $(n-m)$ graus de liberdade ($\mathrm{F}_{\mathrm{in}}$), sendo $n$ o tamanho da amostra e $m$ o número de variáveis no modelo reduzido. Se $\mathrm{F}_{\max} > \mathrm{F}_{\mathrm{in}}$ começa uma nova etapa onde o modelo reduzido terá $m+1$ variáveis, a nova variável que entrará no modelo reduzido será a que possui o $\mathrm{F}_{\max}$. Se $\mathrm{F}_{\max} < \mathrm{F}_{\mathrm{in}}$ interromper o processo e optar pelo modelo reduzido desta etapa.

Também cabe destacar que ao escolher o *Backward Elimination* inicia o processo com o modelo completo e retira uma variável por vez, a cada variável retirada tem-se o modelo reduzido e se calcula a estatística F de cada um deles. Dentre essas a menor $(\mathrm{F_{min}})$ é usada para comparar com o quantil especificado da distribuição F com $1$ e $(n-m-1)$ graus de liberdade $(\mathrm{F_{out}})$, sendo $n$ o tamanho da amostra e $m$ o número de variáveis do modelo completo. Se $\mathrm{F_{min}} < \mathrm{F_{out}}$ uma nova etapa começa, onde o modelo completo terá $m-1$ variáveis, a variável eliminada será a que possui o $\mathrm{F_{min}}$. Se $\mathrm{F_{min}} > \mathrm{F_{out}}$ interromper o processo e optar pelo modelo completo desta etapa.

Igualmente, o *Stepwise Regression* utiliza tanto o *Foward Selection*, quanto o *Backward Elimination* numa mesma etapa, eliminando e adicionado variáveis, para chegar ao melhor modelo.

## 2.5 MULTICOLINEARIDADE

Como exposto por Reynaldo (1997), a multicolinearidade ou mal condicionamento é um grande problema na análise de regressão múltipla, pois ela se refere à situação em que duas ou mais variáveis regressoras de um modelo são moderadamente ou altamente correlacionadas, acarretando uma inflação da variância do estimador dos parâmetros através do método de mínimos quadrados e, possivelmente, dos valores preditos. Além disso, ocorre uma restrição na aplicabilidade do modelo estimado.

Quando se tem o mal condicionamento na matriz das variáveis regressoras, os autovalores da matriz $\boldsymbol{X}^T\boldsymbol{X}$ se aproximam de zero e, dessa forma, a variância do estimador de $\boldsymbol{\beta}$ e o Erro Quadrático Médio Total (EQMT) ficam inflacionados, o que não é conveniente estatisticamente, tendo em vista a necessidade de se minimizar essas quantidades.

Uma forma de detectar multicolinearidade é através da medida conhecida como fator de inflação da variância (VIF), (Berk, 1977), que mede o grau em que cada variável regressora é explicada pelas demais variáveis regressoras, dado por:

$$VIF_i = \frac{1}{1-R_i^2},$$

onde $\mathrm{R_i^2}$ é o coeficiente de determinação da variável regressora $\mathrm{X_i}$ nas demais. Quanto maior o valor de $\mathrm{R_i^2}$ (a variável $\mathrm{X_i}$ é explicada pelas demais variáveis), maior

é o valor de VIF, indicando alta colinearidade. Quando obtemos um VIF de mais de 2,5 começam a indicar níveis relativamente altos de multicolinearidade.

Outra maneira de se identificar a multicolinearidade é através do Índice de Condição Turing (1948), onde são utilizados os autovalores da matriz $\boldsymbol{X}^T\boldsymbol{X}$. Esses autovalores são importantes para a detecção da multicolinearidade, uma vez que quando há colinearidade aproximada entre as regressoras temos autovalores próximos de zero. O índice de condição da matriz $\boldsymbol{X}^T\boldsymbol{X}$ é um conjunto de $p$ valores:

$$\eta_i = \frac{\lambda_{max}}{\lambda_i}, \qquad i = 1,2,\dots,p,$$

onde $\lambda_i$ são os autovalores de $\boldsymbol{X}^T\boldsymbol{X}$. Dessa forma, obtêm-se o número de condição de $\boldsymbol{X}^T\boldsymbol{X}$, que é definido por:

$$\eta = \frac{\lambda_{max}}{\lambda_{min}}.$$

Se o número de condição é menor que 100, não há indícios de um grande problema de mal condicionamento. Entre 100 e 1000 pode se dizer que há um mal condicionamento moderado e para $\eta$ maior que 1000 um mal condicionamento alto.

Convém observar que em aplicações pode se encontrar variáveis regressoras altamente correlacionadas, gerando autovalores da matriz $\boldsymbol{X}^T\boldsymbol{X}$ negativos ou mesmo nulos, impossibilitando o cálculo do índice e do número de condição e até mesmo do VIF.

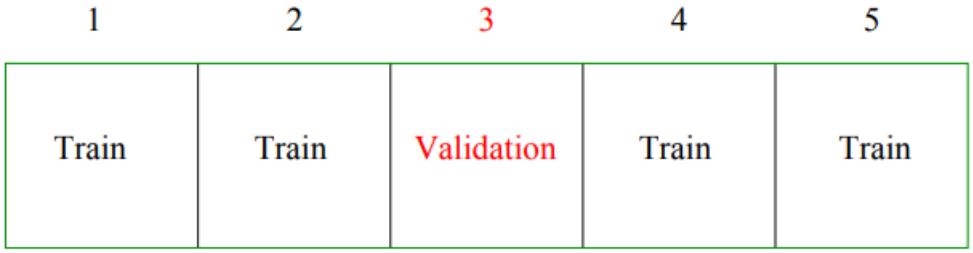
### 2.5.1 VALIDAÇÃO CRUZADA

Para estimar os parâmetros de um modelo onde se tem multicolinearidade, podemos optar por alguns métodos diferentes dos apresentados anteriormente (Mínimos Quadrados e Máxima Verossimilhança). Esses métodos, chamados de métodos de regularização, possui um parâmetro de penalidade $(\lambda)$ desconhecido. Nesse trabalho, utilizaremos o procedimento de validação cruzada para encontrar o melhor valor para $\lambda$.

A Validação Cruzada ou *Cross-Validation* (ver, por exemplo, Hastie, Tibshirani e Friedman, 2009), consiste em um processo que divide o banco de dados em $k$ subconjuntos mutualmente exclusivos, onde $k-1$ subconjuntos serão utilizados como treinamento e 1 subconjunto como validação. O procedimento

consiste em utilizar o subconjunto de treinamento para estimar os parâmetros desejados enquanto o subconjunto de validação é usado para calcular a acurácia (ou erro) desse modelo. Tal processo é repetido $k$ vezes alternando de forma circular um subconjunto de validação com um dos subconjuntos de treinamento a cada iteração. Ao final selecionamos o parâmetro $\lambda$ que retorna o menor erro. A Figura 1 apresenta o esquema de validação cruzada descrito acima.

Figura 1 - Esquema de Validação Cruzada




Fonte: Hastie, Tibshirani e Friedman (2009, p. 242)

Para se fazer a estimação de um bloco qualquer $j$, tem-se que $\boldsymbol{ind_j}$ é o índice das observações bloco $j$. Ainda, seja $\boldsymbol{X_j}^{*T}$ a matriz calculada complementar no $\boldsymbol{ind_j}$ (para o software R, por exemplo, $\boldsymbol{X_j}^{*T} = \boldsymbol{X}[-\boldsymbol{ind_j},]$). Assim, $\widehat{\boldsymbol{\beta}_j}^{*} = (\boldsymbol{X_j}^{*T}\boldsymbol{X_j}^{*})^{-1}\boldsymbol{X_j}^{*T}\boldsymbol{y_j}^{*}$ onde $\boldsymbol{y_j}^{*}$ é o vetor $\boldsymbol{y}$ analisado no complementar de $\boldsymbol{ind_j}$. Portanto, $\widehat{\boldsymbol{y_j}} = \boldsymbol{X_j}\widehat{\boldsymbol{\beta}_j}^{*}$, onde $\boldsymbol{X_j}$ é a matriz $\boldsymbol{X}$ analisada no $\boldsymbol{ind_j}$. Assim, $\widehat{\boldsymbol{y}} = \begin{bmatrix} \widehat{y_1} \\ \vdots \\ \widehat{y_k} \end{bmatrix}$ e sua dimensão é $n\,x\,1$. Desse modo, o valor de $\lambda$ é obtido minimizando a função de Validação Cruzada (CV), dada por:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{1 - h_{ii}}\right)^2,$$

em que $h_{ii}$ é o *i*-ésimo elemento da diagonal de $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

# 3 MÉTODOS DE REGULARIZAÇÃO

Como exposto anteriormente, quando se detecta multicolinearidade nos dados, tem-se a possibilidade de eliminar algumas variáveis, isso pode ser feito através de métodos de seleção de variáveis, introduzidos na Seção 2.4. No entanto, pode ocorrer da variável selecionada para ser excluída não ser a variável com mal condicionamento e então não funcionará para corrigir a multicolinearidade. Outro ponto importante, é que ao se excluir uma variável, temos perca de informação e isso pode levar a uma análise incompleta ou errônea.

Diante de todas essas falhas ao se utilizar métodos de seleção de variáveis para sanar problemas de multicolinearidade, foram criados os métodos de regularização, que serão apresentados a seguir. Esses métodos corrigem a multicolinearidade e apresentam estimadores mais precisos que os métodos usuais.

## 3.1 REGRESSÃO RIDGE

A Regressão Ridge que foi proposta por Hoerl e Kennard (1970) é um método de regularização que tem a finalidade de obtenção de melhores resultados para o modelo, já que a multicolinearidade afeta seriamente as estimativas obtidas pelo método usual de mínimos quadrados. Ela fornece estimativas viciadas, porém de menor variância do que o método usual de mínimos quadrados. A regressão Ridge reduz os coeficientes de regressão impondo uma penalidade em seu tamanho. Essa penalidade é positiva e existem várias formas na literatura para encontrar seu valor ótimo, contudo, nesse trabalho vamos utilizar a validação cruzada, abordada na Seção 2.5.1.

O método Ridge, tal qual o modelo de regressão linear múltipla estimado pelos métodos apresentados na Seção 2.2 e 2.3, permite a interpretação das relações lineares entre as variáveis regressoras e a variável resposta.

### 3.1.1 ESTIMADOR RIDGE

De acordo com Reynaldo (1997, p.36-40) e Pereira (2017, p. 48-53), para tratarem o problema de mal condicionamento, Hoerl e Kennard (1970) definiram um estimador denominado Estimador Ridge utilizando uma penalização no método de quadrados mínimos. Esse estimador obtém uma variância menor que a dos

mínimos quadrados, adicionando uma pequena quantidade positiva, ou seja, viciando o estimador. Ao se permitir viés, o método fornece estimativas melhores no sentido de se diminuir o erro quadrático médio total. A ideia é obter $\boldsymbol{\beta}$ tal que $\mathbf{X\beta}$ esteja o mais próximo possível de $\boldsymbol{y}$. Isso pode ser representado através da seguinte fórmula:

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X\beta}||^2,$$

restrito á elipse imagem da esfera $||\boldsymbol{\beta}||^2 = \mathrm{r}^2$.

A Lagrangeana desse problema é

$$\mathrm{L}(\boldsymbol{\beta}, \lambda) = \left|\left|\mathbf{y} - \mathbf{X\beta}\right|\right|^2 + \lambda(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} - \mathrm{r}^2)$$

$$= (\mathbf{y} - \mathbf{X\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X\beta}) + \lambda(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} - \mathrm{r}^2)$$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathrm{X}\boldsymbol{\beta} - (\mathbf{X\beta})^{\mathrm{T}}\mathbf{y} - (\mathbf{X\beta})^{\mathrm{T}}\mathbf{X\beta} + \lambda(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} - \mathrm{r}^2)$$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X\beta} + \lambda(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} - \mathrm{r}^2)$$

e então,

$$\frac{\partial \mathrm{L}}{\partial \boldsymbol{\beta}} = -2\mathbf{y}^{\mathrm{T}}\mathbf{X} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X\beta} + \lambda(2\boldsymbol{\beta}) = 0$$

$$= -\mathbf{y}^{\mathrm{T}}\mathbf{X} + (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta} = 0$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}}(\mathrm{r}^2) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

No espaço de dados, tem-se que:

$$||\,\mathbf{y} - \mathbf{X\beta}||^2 = ||\,\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}||^2 + ||\,\mathbf{X\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}||^2 = \mathrm{cte} + ||\,\mathbf{X\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}||^2,$$

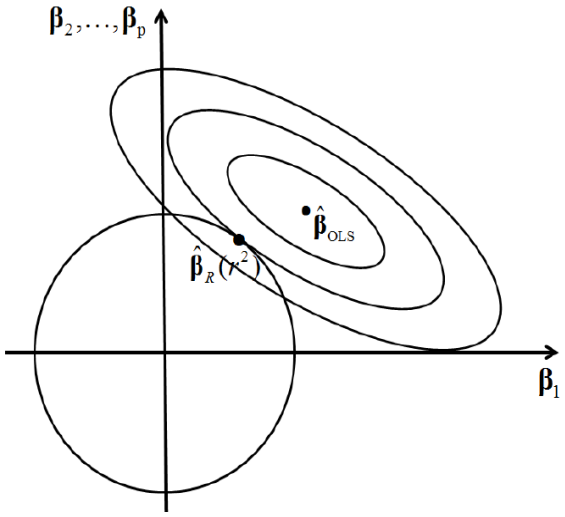
em que $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$ é o estimador de mínimos quadrados.

O que se quer então é a minimização

$$\min_{\boldsymbol{\beta}} ||\,\mathbf{X\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2 = \min_{\boldsymbol{\beta}} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right),$$

sujeito a restrição $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = \mathrm{r}^2$, isto é, para os vetores $\boldsymbol{\beta}$ sobre a esfera, obter aquele na elipse centrada em $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$ de menor tamanho pela métrica de Mahalanobis, como na Figura 2.

Figura 2 - Projeção definida pela métrica de Mahalanobis




Fonte: Pereira (2017, p. 53)

Como é possível obter a forma analítica do estimador ridge, podemos obter o seu Erro Quadrático Médio (EQM):

$$\begin{aligned}
\mathrm{EQM}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}, \boldsymbol{\beta}\right) &= \mathrm{E}\left\{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{R}} - \boldsymbol{\beta}\right)^{\mathrm{T}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{R}} - \boldsymbol{\beta}\right)\right\} \\
&= \mathrm{E}\left\{\widehat{\boldsymbol{\beta}}_{\mathrm{R}}{}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{R}} - \widehat{\boldsymbol{\beta}}_{\mathrm{R}}{}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{R}} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\} \\
&= \sigma^2\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \\
&= \sigma^2\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) + \boldsymbol{\beta}^{\mathrm{T}}\left[\mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\mathbf{X} + \mathbf{I}_{\mathrm{p}}\right]\boldsymbol{\beta},
\end{aligned}$$

onde $\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_{\mathrm{p}})^{-1}\mathbf{X}^{\mathrm{T}}$ e $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta}$.

Conforme pode ser notado na Figura 2, dificilmente o método Ridge zera um coeficiente de regressão. Portanto, se o objetivo do pesquisador é eliminar o problema de multicolinearidade e ao mesmo tempo manter todas as variáveis no modelo, esse método é uma boa solução.

## 3.2 LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR - LASSO

Ainda abordado por Pereira (2017), além de se fazer necessário um robusto método de estimação dos parâmetros, pode surgir também o interesse em selecionar algumas variáveis do modelo. Se ao utilizar o estimador de mínimos quadrados o mesmo apresentar alta variabilidade, devido à alta correlação entre as variáveis regressoras (multicolinearidade), essa estimação não será tão precisa, ou seja, se um novo vetor de respostas $\boldsymbol{Y}$ é observado, há grandes chances de variáveis regressoras diferentes serem selecionadas.

Foi então proposto por Tibshirani (1996) outro método de regularização, denominado LASSO (Least Absolute Shrinkage and Selection Operator), esse método tem por objetivo estimar os coeficientes de regressão ($\boldsymbol{\beta's}$), através de uma penalização, que além de gerar um modelo de regressão linear que apresenta pouca variabilidade, gere também estimativas de $\widehat{\boldsymbol{\beta}}$ em que várias de suas componentes sejam nulas, e, portanto, o método é também um método automático de seleção de variáveis. É válido ressaltar que esse método seleciona no máximo o número de variáveis igual ao número de observações.

O método LASSO, tal qual o modelo de regressão linear múltipla estimado pelos métodos apresentados na Seção 2.2 e 2.3 e o método Ridge, permite a interpretação das relações lineares entre as variáveis regressoras e a variável resposta.

### 3.2.1 ESTIMADOR LASSO

O estimador LASSO penaliza o método de mínimos quadrados, tornando o estimador viciado. O estimador pode ser interpretado como:

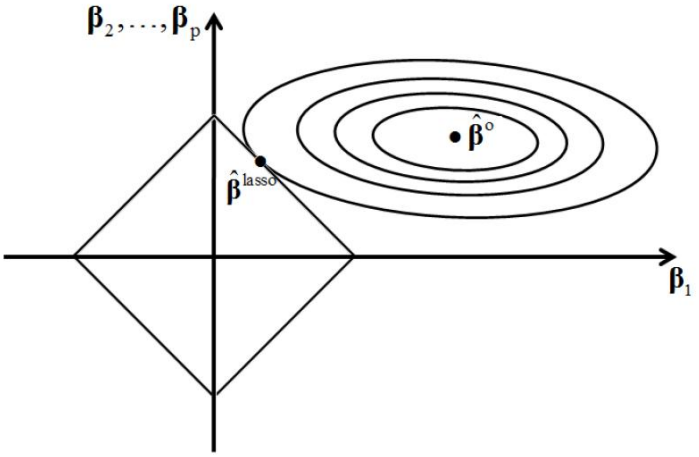
$$\operatorname{argmin}\left|\left|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right|\right|^{2},$$

sujeito a restrição $\sum_{j=1}^{p}|\beta_j| \leq t$.

Essa última forma fica representada na forma Lagrangeana por:

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\beta\in\mathbb{R}^p}\left(||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||^2 + \lambda\left|\left|\boldsymbol{\beta}\right|\right|_1\right.$$

Figura 3 – Obtenção do Estimador LASSO




Fonte: Pereira (2017, p. 56)

A Figura 3 representa a obtenção do estimador LASSO. O desenvolvimento analítico desse estimador é bastante complexo, então se faz necessário utilizar métodos computacionais. Para isso, o software R Core Team (2018) apresenta pacotes que programam o cálculo de estimativas dos coeficientes de regressão utilizando o método LASSO. O pacote utilizado nesse trabalho será o “glmnet”, desenvolvido por Friedman, Hastie e Tibshirani (2010).

## 3.3 MÉTODOS DE REDUÇÃO DE DIMENSIONALIDADE

Conforme Hastie, Tibshirani e Friedman (2009, p. 79-82), esses métodos projetam os $p$ preditores a um subespaço M-dimensional, $M < p$. Os métodos têm por objetivo calcular $M$ diferentes combinações lineares, ou projeções, desses preditores. Então essas $M$ projeções são usadas como novos preditores da variável resposta. Neste trabalho serão abordados dois métodos de redução de dimensão, PCR e PLS, os mesmos diferem em como as combinações lineares são construídas.

Convém notar que nos métodos de redução de dimensionalidade as variáveis são transformadas. O objetivo desses métodos é fazer predição, não havendo interesse na interpretação das relações lineares entre as variáveis explicativas e a variável resposta.

### 3.3.1 REGRESSÃO DE COMPONENTES PRINCIPAIS- PCR

É um método que tem por finalidade básica a análise dos dados usados visando sua redução, eliminação de sobreposições e a escolha das formas mais representativas de dados a partir de combinações lineares das variáveis originais. O método PCR é bastante utilizado quando os vetores de características têm muitas dimensões. Considere $\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_M}$ as $M < p$ combinações lineares das $p$ variáveis regressoras originais, isto é:

$$\boldsymbol{Z_m} = \sum_{j=1}^{p} \emptyset_{jm} \boldsymbol{X_j}$$

para algumas constantes $\emptyset_{1m}, \emptyset_{2m}, \ldots, \emptyset_{pm}$, $m = 1, \ldots, M$. Para a escolha de $M$ e a estimação dos $\emptyset_{ij}$ ver, por exemplo, Johnson e Wichern (2007). Portanto, o novo modelo de regressão linear é dado por:

$$Y_i = \theta_0 + \theta_1 Z_{i1} + \cdots + \theta_M Z_{iM} + \epsilon_i, \qquad i = 1, \ldots, n.$$

Logo,

$$\widehat{\boldsymbol{\theta}}^{\mathrm{pcr}}(\mathrm{M}) = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{Y},$$

sendo $\boldsymbol{Z} = [\mathbf{1}\ \boldsymbol{Z_1}\ \boldsymbol{Z_2}\ \ldots\ \boldsymbol{Z_M}]$.

### 3.3.2 MÍNIMOS QUADRADOS PARCIAIS- PLS

Essa técnica também constrói um conjunto de combinações lineares das entradas para a regressão, mas ao contrário da PCR, usa $\boldsymbol{Y}$ (além de $\boldsymbol{X}$) para essa construção. Como a regressão de componentes principais o PLS também não é invariante de escala, portanto, assumimos que cada $\mathbf{X_j}$ é padronizado, produzindo $\boldsymbol{X}^{(\mathbf{0})}_{\mathbf{j}}$, tal que, $\boldsymbol{X}^{(\mathbf{0})}_{\mathbf{ij}} = \frac{\boldsymbol{X_{ij}} - \overline{\boldsymbol{X_j}}}{\boldsymbol{sd}(\boldsymbol{X_j})}$.

O PLS começa pela computação $\emptyset_{j1}$, onde $\emptyset_{j1}$ é o coeficiente de inclinação da regressão linear simples de $\boldsymbol{Y}$ em $\boldsymbol{X}^{(\mathbf{0})}_{\mathbf{j}}$ para cada $j$, $j = 1, \ldots, p$. A partir disso, construímos a entrada derivada $\mathbf{Z_1} = \sum_{\mathbf{j}=1}^{p} \emptyset_{j1} \boldsymbol{X}^{(\mathbf{0})}_{\mathbf{j}}$ que é a primeira direção parcial dos mínimos quadrados. Ortogonalizamos $\boldsymbol{X}^{(\mathbf{0})}_{\mathbf{1}}, \ldots, \boldsymbol{X}^{(\mathbf{0})}_{\mathbf{p}}$ em relação à $\mathbf{Z_1}$, obtendo

$\boldsymbol{X^{(1)}_j} = \boldsymbol{X^{(0)}_j} - [< \mathbf{Z_1}, \boldsymbol{X^{(0)}_j} > / < \mathbf{Z_1}, \mathbf{Z_1} >] \mathbf{Z_1},\ j = 1, \ldots, p$ (por exemplo, $< \boldsymbol{x}, \boldsymbol{y} >$ é o produto interno de $\boldsymbol{x}$ e $\boldsymbol{y}$). Tem-se que $\mathbf{Z_2} = \sum_{j=1}^{p} \emptyset_{j2}\, \boldsymbol{X^{(1)}_j}$, onde $\emptyset_{j2}$ é o coeficiente de inclinação da regressão linear simples de $\boldsymbol{Y}$ em $\boldsymbol{X^{(1)}_j}$.

Continuamos esse processo até que $M \leq p$ direções tenham sido obtidas, para o algoritmo iterativo completo, ver Hastie, Tibshirani e Friedman (2009, p. 81).

Portanto, o novo modelo de regressão linear é dado por:

$$Y_i = \theta_0 + \theta_1 Z_{i1} + \cdots + \theta_M Z_{iM} + \epsilon_i, \qquad i = 1, \ldots, n.$$

Logo,

$$\widehat{\boldsymbol{\theta}}^{\mathrm{pls}}(\mathrm{M}) = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{Y},$$

sendo $\boldsymbol{Z} = [\mathbf{1}\ \boldsymbol{Z_1}\ \boldsymbol{Z_2}\ \ldots\ \boldsymbol{Z_m}]$.

O PLS busca direções que tem alta variância e tem alta correlação com a resposta, em contraste com PCR, que só ocorre em alta variância.

Apesar de esse trabalho apresentar os dois métodos, na parte de aplicação será trabalhado o método de mínimos quadrados parciais.

# 4 APLICAÇÕES

Neste capítulo serão realizadas aplicações em dados reais para os métodos apresentados nas seções anteriores. Todo o processo será feito no software R Team (2018), utilizando pacotes já disponíveis em sua biblioteca e também algoritmos desenvolvidos durante esse trabalho.

## 4.1 BANCO DE DADOS - BAILARINAS

Para ilustrar o método de seleção de variáveis será utilizado um banco de dados disponível no software R - *CharnetApD.1*. Esse banco contém 163 observações de medidas de ângulos de rotação dos pés de meninas bailarinas como variável resposta, e, idade, peso e altura como variáveis regressoras.

Será utilizada a função *vif* do pacote *car* (John FoxSanford e Weisberg, 2011) para verificar se existe multicolinearidade entre as variáveis regressoras. Em seguida, será feito seleção de variáveis através do método *Backward Elimination* e os três critérios apresentados anteriormente serão aplicados. Para os critérios AIC e BIC utilizou-se a função *stepwise* do pacote *RcmdrMisc* (Fox, 2018). Para calcular a Estatística F foi desenvolvido um algoritmo durante esse trabalho.

Tabela 1 - Valores de VIF para os dados da bailarina.

| **VIF** | | |
|---|---|---|
| **Idade** | **Peso** | **Altura** |
| 5,241 | 6,362 | 9,565 |

Tabela 2 - Estimativas dos parâmetros (erros-padrão entre parênteses) para o modelo a partir do método *Backward* em diferentes critérios.

| **Critérios** | Intercepto | Altura | Peso | Idade |
|---|---|---|---|---|
| **AIC** | -0,513(13,021) | 0,289(0,139) | -0,249(0,159) | 2,195(0,519) |
| **BIC** | 25,840(2,464) | - | - | 2,748(0,228) |
| **Estatística F** | 25,840(2,464) | - | - | 2,748(0,228) |

A Tabela 1 apresenta todos os valores de VIF maiores que 2,5 e pela Seção 2.5, conclui-se então, que os dados apresentam multicolinearidade. Diante disso,

pode se considerar a eliminação de algumas variáveis do modelo. Para isso foi utilizado o método *Backward Elimination* e os resultados foram apresentados na Tabela 2. Os critérios BIC e Estatística F são mais rigorosos e eliminaram duas variáveis: altura e peso, enquanto o AIC não eliminou nenhuma variável. Considerando um modelo com a variável idade e o intercepto, o mesmo teria apenas uma variável regressora e, consequentemente, a multicolinearidade seria eliminada. Entretanto, informações de duas variáveis seriam perdidas.

Portanto, o modelo final selecionado é $y = 25{,}840 + 2{,}748\text{Idade}$, onde $y$ é a medida do ângulo de rotação dos pés de meninas bailarinas. Assim, a cada ano a mais aumenta-se em média 2,75 graus o ângulo de rotação dos pés.

## 4.2 BANCO DE DADOS - PRÓSTATA

A segunda aplicação consiste em um estudo que examinou a correlação entre o nível de antígeno específico da próstata e uma série de medidas clínicas em homens que estavam prestes a receber uma prostatectomia radical. É uma base de dados do R denominada *Prostate*, com 97 linhas e 8 variáveis regressoras. Algumas variáveis desse banco apresentam alta correlação, então os métodos de regularização se mostram eficientes. Para ilustrar esses métodos serão utilizadas as funções *cv.glmnet* (uma função do pacote *glmnet*), que encontra o parâmetro de penalidade $\lambda$ através de validação cruzada para os métodos Ridge e LASSO, e *plsr* que faz parte do pacote *pls* (Mevik, Wehrens e Liland, 2016) para o método PLS. Os dados serão divididos em duas partes: treino (67%) e validação (33%).

Tabela 3 - Valores de VIF para Prostate.

| VIF | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Lweight** | **Age** | **Lbph** | **Svi** | **Lcp** | **Gleason** | **Pgg45** | **Lpsa** |
| 1,471 | 1,306 | 1,392 | 2,167 | 2,558 | 2,433 | 2,919 | 1,922 |

Analisando a Tabela 3, as variáveis regressoras Lcp e Pgg45 apresentam multicolinearidade. Agora que já foi verificado a multicolinearidade, os resultados dos métodos serão apresentados.

É possível obter uma gama de valores para $\lambda$, como apresentado nas Figuras 4 e 6, e de acordo com que essa penalidade varia, os valores dos coeficientes

mudam. Utilizando a função *cv.glmnet* podemos obter $\lambda's$ "ótimos" para Ridge e LASSO. As Figuras 5 e 7 apresentam a variação do Erro Quadrático Médio (EQM) em função do logaritmo de $\lambda$. No Ridge, sempre vamos manter com oito variáveis e no Lasso algumas variáveis serão excluídas.

Pela Figura 4, de acordo que log($\lambda$) aumenta os coeficientes vão ficando mais próximos de 0, em contrapartida tem-se pela Figura 5 que de acordo com que log($\lambda$) aumenta, o EQM também aumenta. Por esse motivo calcula-se o $\lambda$ que produz menor EQM ($\lambda$ mínimo) e calcula $\lambda$ mais o desvio padrão (esse $\lambda$ é calculado, pois no LASSO ele consegue eliminar mais variável no modelo e continua com um EQM próximo do EQM mínimo), obtendo assim um intervalo de $\lambda$ "ótimos", como mencionado acima e apresentado nas linhas pontilhadas dos gráficos abaixo.

Figura 4 - Variação dos coeficientes Ridge de acordo com aumento de log λ.

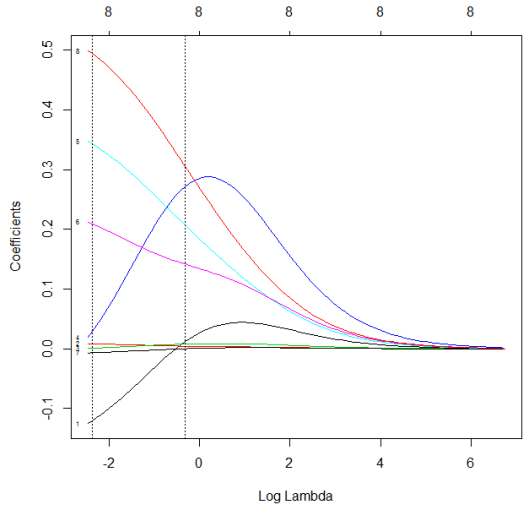
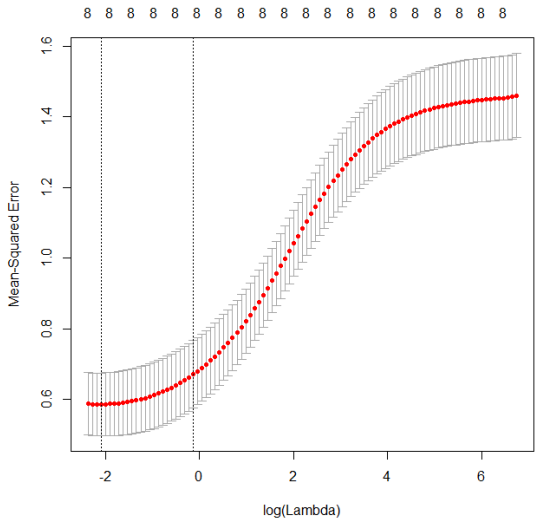



Figura 5 - Gráfico da variação do log de $\lambda$ de acordo com EQM – Ridge.



Na Figura 5 tem-se que os melhores valores para log($\lambda$) estão entre -2 e 0 ($\lambda$ entre 0,14 e 1) e o modelo mantém as oito variáveis, uma vez que o método Ridge não atua como selecionador de variável.

Da mesma forma, pela Figura 6 tem-se que com o aumento de log $\lambda$ os coeficientes tendem a 0, entretanto no método LASSO os coeficientes zeram e, consequentemente, o mesmo atua como um método de seleção de variáveis.

Figura 6 - Variação dos coeficientes LASSO de acordo com aumento de log $\boldsymbol{\lambda}$.

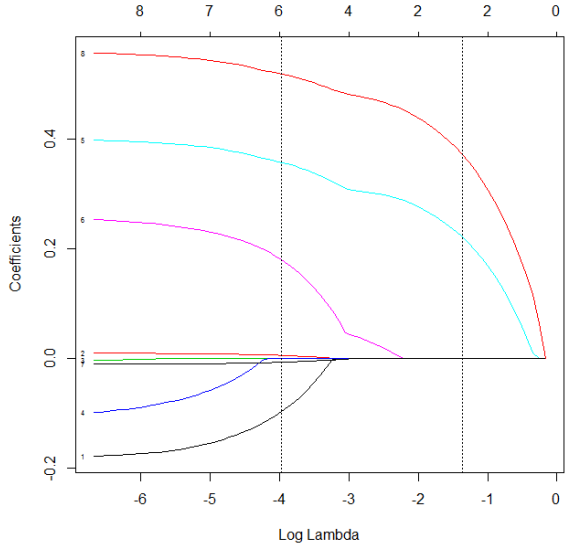
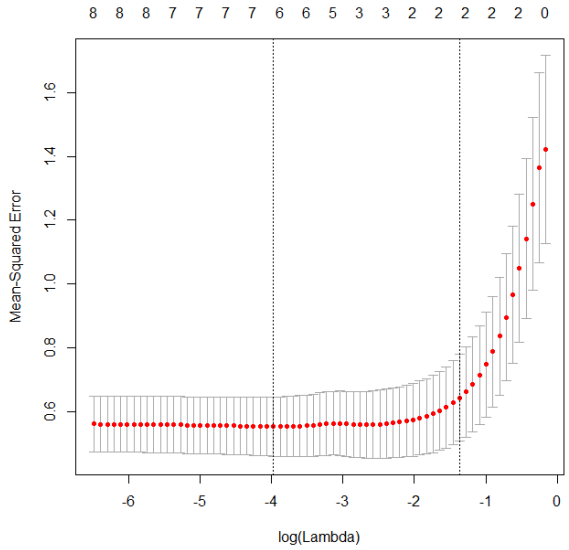


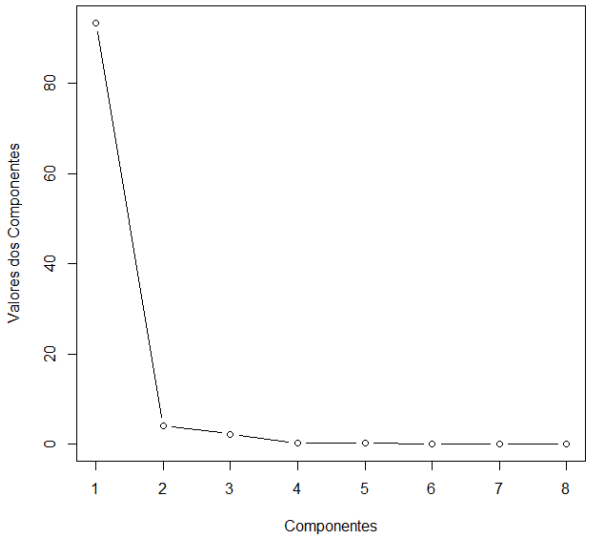

Figura 7 - Gráfico da variação do log de $\boldsymbol{\lambda}$ de acordo com EQM – LASSO.



Na Figura 7 tem-se que os melhores valores para log(λ) estão entre -4 e -1,8 ($\lambda$ entre 0,02 e 0,17) e dependendo do λ escolhido, o modelo pode ser reduzido a 6, 5, 3 ou 2 variáveis.

Na Figura 8 são apresentados os valores das componentes para o método PLS, se o número de componentes escolhido for oito, tem se o modelo estimado pelo método de mínimos quadrados e não se tem nenhuma redução de dimensionalidade. Aqui serão escolhidas três componentes, pois é perceptível uma estabilidade a partir desse número.

Figura 8 - Gráfico da distribuição dos valores das componentes - PLS.




A Tabela 4 apresenta o número de variáveis e valor do EQM do modelo final para cada método. O pior desempenho fica com o método PLS, pois sabe se que esse método se adapta muito bem onde todas as variáveis do modelo tem alta correlação, como em dados de NIR. O que não é o caso desse banco de dados, pois aqui nem todas as variáveis são mal condicionadas. Os métodos de Mínimos Quadrados, Ridge e LASSO apresentam EQM's muito próximos.

Tabela 4- EQM's e número de variáveis do modelo para os métodos MQ, Ridge, LASSO e PLS.

| | **MMQ** | **Ridge** | **LASSO** | **PLS** |
|---|---|---|---|---|
| **EQM** | 0,492 | 0,484 | 0,485 | 0,514 |
| **Num. Variável** | 8 | 8 | 6 | 3 |

A Tabela 5 apresenta as estimativas dos parâmetros do modelo. Para a estimação de Ridge e LASSO foi utilizado o $\lambda$ mínimo. O método PLS não é apresentado aqui, pois o método perde a interpretação das variáveis originais.

Tabela 5 - Estimação dos parâmetros e EQM para MMQ, Ridge e LASSO.

| **Métodos** | **Variáveis** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Lweight** | **Age** | **Lbph** | **Svi** | **Lcp** | **Gleason** | **Pgg45** | **Lpsa** |
| **MMQ** | -0,182 | 0,011 | -0,004 | -0,109 | 0,402 | 0,260 | -0,010 | 0,561 |
| **Ridge** | -0,111 | 0,008 | 0,002 | 0,049 | 0,334 | 0,203 | -0,006 | 0,483 |
| **LASSO** | -0,097 | 0,006 | - | - | 0,358 | 0,181 | -0,006 | 0,519 |

Como esperado, os valores da estimação dos parâmetros diminuíram em módulo na regressão Ridge em relação ao método de MQ e no LASSO além de diminuir os valores dos parâmetros em módulo, duas variáveis foram eliminadas.

Focando no objetivo de selecionar o modelo mais parcimonioso, podemos escolher o método LASSO, pois elimina duas variáveis do modelo e apresenta um bom EQM.

### 4.3 BANCO DE DADOS - GASOLINA

Nesta seção será apresentado o banco de dados *Gasoline*, também disponível na biblioteca do R. Esse banco consiste em um conjunto de dados com espectros NIR e números de octanas de 60 amostras de gasolina. Os espectros NIR foram medidos usando refletância difusa de 900 nm a 1700 nm em intervalos de 2 nm, totalizando em 401 comprimentos de onda. Quando se tem como variáveis explicativas espectros de NIR, o método de mínimos quadrados parciais (PLS) se mostra muito eficiente, por se tratar de uma mesma variável sendo aplicada em diferentes intensidades. Neste caso é fácil notar que as variáveis regressoras apresentam multicolinearidade, mas uma medida apropriada para se calcular esse mal condicionamento é o número de condição, apresentado na seção 2.5, uma vez que o VIF retornaria 401 valores para serem analisados. Entretanto, para encontrar essa medida precisamos dos autovalores da matriz $\boldsymbol{X^T X}$ e como os dados tem uma

alta correlação, acarreta em uma matriz singular, ou seja, não inversível, trazendo autovalores negativos e muito próximos de zero, impossibilitando o cálculo do número de condição.

Dessa forma, vamos partir do pressuposto que existe multicolinearidade nos dados e aplicar os métodos de regularização através dos pacotes já citados acima. Nos métodos Ridge e LASSO foi utilizado o $\lambda$ mínimo para a obtenção dos coeficientes. No método PLS, para a escolha do número de componentes ideal, analisou-se a distribuição dos valores e verificou onde a mesma se estabilizou. Os dados também serão divididos em duas partes: treino (67%) e validação (33%).

Pela Figura 9, percebe-se que de acordo com que o número de componentes vai aumentando o valor da estimação dos componentes vai ficando próximo de zero. Assim, optou-se por utilizar 6 componentes.

Tabela 6 - EQM's e número de variáveis do modelo para os métodos Ridge, LASSO e PLS.

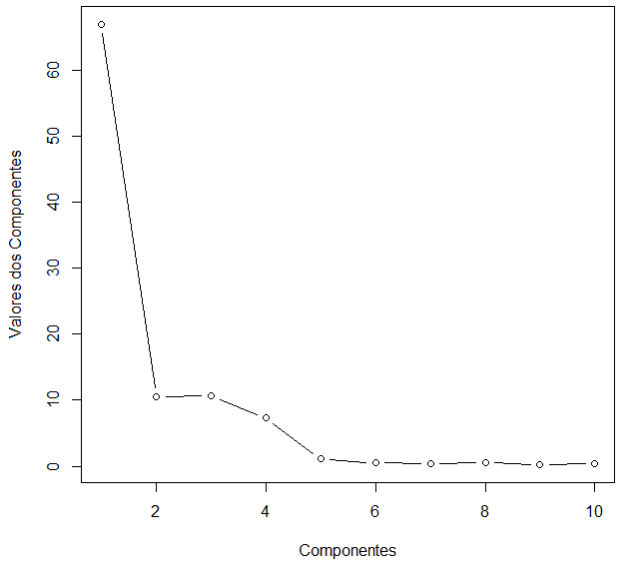
| | **Ridge** | **LASSO** | **PLS** |
|---|---|---|---|
| **EQM** | 0,365 | 0,058 | 0,059 |
| **Num. Variável** | 401 | 12 | 6 |

Através da Tabela 6, percebe se que os métodos LASSO e PLS são bons métodos para se utilizar quando se tem dados de NIR. Os EQM's ficam bem próximos e como o interesse é na predição, pode se optar pelo método PLS, cujo modelo final contém apenas 6 componentes e que se ajusta bem aos dados. Na Tabela 7 apresentamos os valores das componentes.

Tabela 7 - Valores da estimação dos 6 componentes através do método PLS.

| **Componentes** | | | | | |
|---|---|---|---|---|---|
| **Comp1** | **Comp2** | **Comp3** | **Comp4** | **Comp5** | **Comp6** |
| 66,918 | 10,487 | 10,675 | 7,328 | 1,076 | 0,514 |

Figura 9 - Gráfico da distribuição dos valores das componentes.

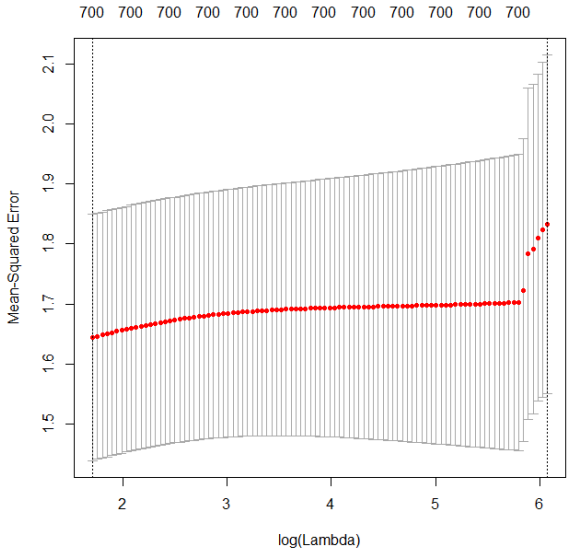



## 4.4 BANCO DE DADOS – CELULOSE

Por fim, serão aplicados os métodos Ridge, LASSO e PLS em uma base de dados de uma empresa de produção de celulose. A base possui 700 variáveis regressoras, cada uma dessas variáveis representa uma intensidade de feixe de luz que são aplicados na madeira da árvore para obter a absorbância da mesma, semelhante aos dados de NIR expostos na aplicação anterior. Foram analisadas 118 árvores e tem se como variável resposta o rendimento da madeira. Nesta aplicação os dados também serão divididos em duas partes: treino (67%) e validação (33%) e as funções *cv.glmnet* e *plsr* serão utilizadas.

Pelos mesmos motivos do banco de dados anterior, não é possível calcular as medidas para detectar o mal condicionamento, mas é nítido que as variáveis regressoras possuem mal condicionamento, uma vez que é a mesma variável sendo medida em diferentes níveis.

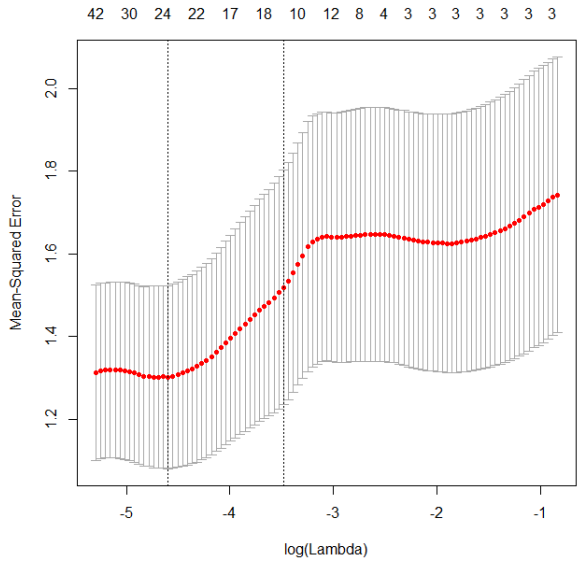
Figura 10 - Gráfico da variação do log de λ de acordo com EQM – Ridge.




Para o método Ridge, a Figura 10 mostra que apesar da variação log($\lambda$), o EQM não varia muito. E como esperado o modelo mantém as 700 variáveis, pois o método Ridge não seleciona variáveis.

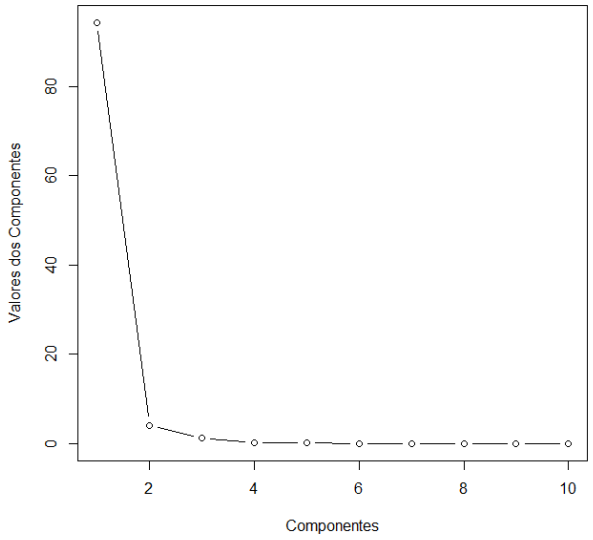
Para o método LASSO, a Figura 11 mostra que os melhores valores para log(λ) estão entre -4,5 e -3,5 ($\lambda$ entre 0,01 e 0,03) e dependendo do valor de λ escolhido nesse intervalo, pode se ter um modelo de 15 a 26 variáveis, uma vez que o método LASSO também funciona como um método de seleção de variáveis.

Figura 11 - Gráfico da variação do log de λ de acordo com EQM – LASSO.



Para o método PLS, analisando a Figura 12, a estabilização ocorre a partir de 4 componentes.

Figura 12 - Gráfico da distribuição dos valores das componentes.




Na estimação dos parâmetros, foi utilizado o $\lambda$ mínimo para os métodos Ridge e LASSO e utilizaram-se quatro componentes para o método PLS. A Tabela 8 apresenta os EQM's para os três métodos e o número de variáveis selecionadas para compor o modelo.

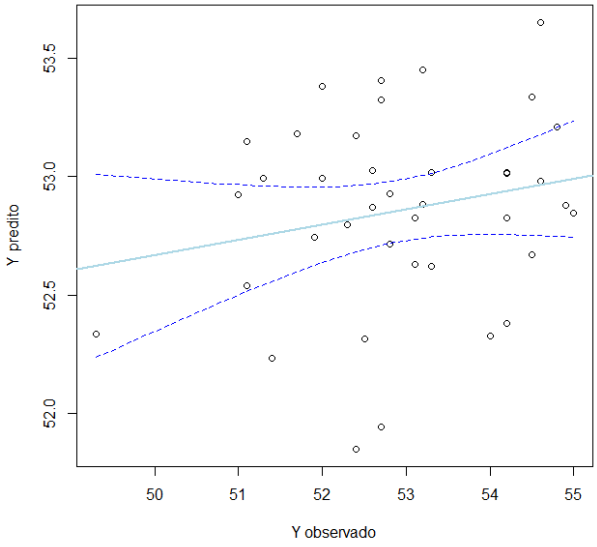
Tabela 8 - EQM's e número de variáveis do modelo para os métodos Ridge, LASSO e PLS.

| | **Ridge** | **LASSO** | **PLS** |
|---|---|---|---|
| **EQM** | 1.615 | 0.963 | 0.914 |
| **Num. Variável** | 700 | 26 | 4 |

Através dos EQM's da Tabela 8, temos como pior modelo o estimado pelo método Ridge. Os modelos LASSO e PLS apresentam EQM's próximos, ainda sim o PLS tem o melhor ajuste aos dados.

A intenção nesse banco de dados não é a interpretação das variáveis regressoras, o interesse é apenas na predição, então o método PLS será a melhor opção aqui. O modelo através desse método se ajusta bem aos dados, e ainda reduz de 700 variáveis regressoras para 4 componentes.

Figura 13 - Y observado versus Y predito - Método Ridge. Reta ajustada com bandas de confiança.




Em linha com o EQM, o gráfico do valor de $y$ observado versus o $y$ predito para o método Ridge na Figura 13, mostrou que o modelo não se ajustou bem aos dados.

Figura 14 - Y observado versus Y predito - Método LASSO. Reta ajustada com bandas de confiança.

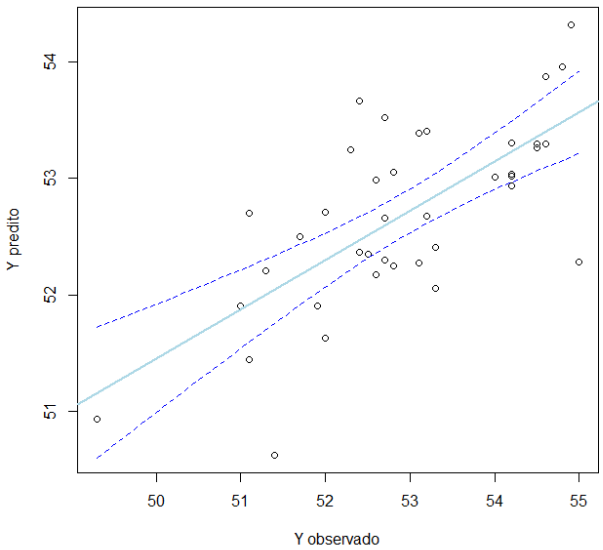
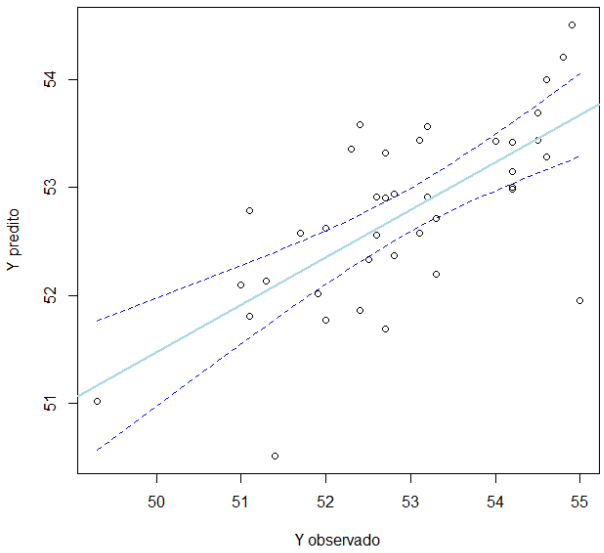


Figura 15 - Y observado versus Y predito - Método PLS. Reta ajustada com bandas de confiança




Para os gráficos das Figuras 14 e 15, nos métodos LASSO e PLS é notável uma variabilidade menor, os modelos se ajustam melhor aos dados. Como não se tem o objetivo apenas de predição, podemos optar pelo método PLS, com menor número de parâmetros para o modelo final e baixo custo computacional.

## 5 CONSIDERAÇÕES FINAIS

No estudo de seleção de modelos de regressão linear múltipla é comum se deparar com vários métodos, cada um com suas particularidades. Ao estudar e analisar vários deles, como apresentado acima, percebe-se que não existe um método que seja melhor que o outro em todos os aspectos. É necessário analisar o banco de dados em que se vai trabalhar e saber quais os objetivos finais. Se o objetivo for apenas diminuir o número de variáveis regressoras, talvez seja interessante utilizar os métodos de seleção de variáveis: *Backward*, *Forward* ou *Stepwise*. Caso seja apenas a regularização do banco de dados, corrigindo a multicolinearidade, poderia se pensar em utilizar a Regressão Ridge. E também temos o LASSO e PLS que além corrigir a multicolinearidade, funcionam como selecionador de variáveis regressoras. As aplicações foram de extrema importância para compreender melhor cada um dos métodos.

## REFERÊNCIAS


AKAIKE, H. **A new look at the statistical model identification**. *IEEE Transactions on Automatic Control*, 19(6): p.716-723, 1974.

BERK, Kenneth N. Tolerance and condition in regression computations. **Journal of the American Statistical Association**, v. 72, n. 360a, p. 863-866, 1977.

BOZDOGAN, Hamparsum. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. **Psychometrika**, v. 52, n. 3, p. 345-370, 1987.

CHARNET, Reinaldo et al. **Análise de modelos de regressão linear com aplicações**. Campinas, São Paulo, Unicamp, 356p, 1999.

GALTON, Francis. **Natural inheritance**. New York: AMS Press, 1889.

DI GANGI, Leonardo et al. **An efficient optimization approach for best subset selection in linear regression, with application to model selection and fitting in autoregressive time-series**. Computational Optimization and Applications, v. 74, n. 3, p. 919-948, 2019.

FOX, John (2018). **RcmdrMisc: R Commander Miscellaneous Functions**. R package version 1.0-10. Disponível em <https://CRAN.R-project.org/package=RcmdrMisc>.

FOX, John and WEISBERG, Sanford (2011). **An {R} Companion to Applied Regression**, Second Edition. Thousand Oaks CA: Sage. Disponível em: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>.

HAALAND, David M.; THOMAS, Edward V. Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. **Analytical Chemistry**, v. 60, n. 11, p. 1202-1208, 1988.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2nd ed. New York, NY: Springer, 2009. (Springer series in statistics).

HOERL, Arthur E.; KENNARD, Robert W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, v. 12, n. 1, p. 55–67, 1970.

MATEOS-APARICIO, Gregoria. **Partial least squares (PLS) methods: Origins, evolution, and application to social sciences**. Communications in Statistics-Theory and Methods, v. 40, n. 13, p. 2305-2317, 2011.

MELKUMOVA, L. E.; SHATSKIKH, S. Ya. **Comparing Ridge and LASSO estimators for data analysis**. Procedia engineering, v. 201, p. 746-755, 2017.

MEVIK, Bjørn-Helge; WEHRENS, Ron and LILAND, Kristian Hovde (2016). pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0. Disponível em <https://CRAN.R-project.org/package=pls>.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to linear regression analysis**. 5th ed. Hoboken, NJ: Wiley, 2012. (Wiley series in probability and statistics, 821).

PEREIRA, Leandro da Silva. **Geometria dos métodos de regressão LARS, LASSO e Elastic Net com uma aplicação em seleção genômica**. p. 167. Tese (Doutorado) - Curso de Estatística, Universidade Federal de Lavras, Lavras,Mg, 2017.

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

REYNALDO, Cristiane. **Um método alternativo para o Mal Condicionamento da Matriz de Regressoras**. Dissertação de Mestrado, Universidade Estadual de Campinas, Campinas, Sp, 1997.

RICHARD, A. Johnson; Dean, W. Wichern. **Applied multivariate statistical analysis**. 2007. 6nd ed. Upper Saddle River. Pearson Prentice Hall.

RODRIGUES, Sandra. **Modelo de Regressão Linear e suas Aplicações**. Tese (Mestrado) - Universidade da Beira Interior, Covilhã, Portugal, 2012.

SCHWARZ, Gideon. Estimating the Dimension of a Model. **Annals of Statistics** 6: 461-64, 1978.

SIMON, Noah; FRIEDMAN, Jerome; HASTIE, Trevor; Tibshirani, Rob (2011). **Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent**. Journal of Statistical Software, 39(5), 1-13. Disponível em: <http://www.jstatsoft.org/v39/i05/>.

TIBSHIRANI, Robert. Regression Shrinkage and Selection Via the Lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

TURING, Alan Mathison. **Intelligent machinery**. 1948.