Comparação de métodos de estimação para problemas com colinearidade e/ou alta dimensionalidade (p > n)

Marcelo Henrique Casagrande

Marcelo Henrique Casagrande

Comparação de métodos de estimação para problemas com colinearidade e/ou alta dimensionalidade (p > n)

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
Junho de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

C334c

Casagrande, Marcelo
Comparação de métodos de estimação para problemas com colinearidade e/ou alta dimensionalidade (p > n) / Marcelo Casagrande; orientador Carlos Diniz. -- São Carlos, 2016.
65 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. regressão ridge. 2. LASSO. 3. mínimos quadrados parciais. 4. regressão por componentes principais. 5. alta dimensionalidade. I. Diniz, Carlos, orient. II. Título.

Marcelo Henrique Casagrande

Comparison of estimation methods for problems with collinear and/or high dimensionality (p > n)

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP and to the Departamento de Estatística – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
June 2016

Resumo

Este trabalho apresenta um estudo comparativo do poder de predição de quatro métodos de regressão adequados para situações nas quais os dados, dispostos na matriz de planejamento, apresentam sérios problemas de multicolinearidade e/ou de alta dimensionalidade, em que o número de covariáveis é maior do que o número de observações.

No presente trabalho, os métodos abordados são: regressão por componentes principais, regressão por mínimos quadrados parciais, regressão ridge e LASSO.

O trabalho engloba simulações, em que o poder preditivo de cada uma das técnicas é avaliado para diferentes cenários definidos por número de covariáveis, tamanho de amostra e quantidade e intensidade de coeficientes (efeitos) significativos, destacando as principais diferenças entre os métodos e possibilitando a criação de um guia para que o usuário possa escolher qual metodologia usar com base em algum conhecimento prévio que o mesmo possa ter.

Uma aplicação em dados reais (não simulados) também é abordada.

Palavras chaves: regressão ridge, LASSO, mínimos quadrados parciais, regressão por componentes principais, alta dimensionalidade

Abstract

This paper presents a comparative study of the predictive power of four suitable regression methods for situations in which data, arranged in the planning matrix, are very poorly multicolinearity and / or high dimensionality, wherein the number of covariates is greater the number of observations.

In this study, the methods discussed are: principal component regression, partial least squares regression, ridge regression and LASSO.

The work includes simulations, wherein the predictive power of each of the techniques is evaluated for different scenarios defined by the number of covariates, sample size and quantity and intensity ratios (effects) significant, highlighting the main differences between the methods and allowing for the creating a guide for the user to choose which method to use based on some prior knowledge that it may have.

An application on real data (not simulated) is also addressed.

Key words: ridge regression, LASSO, partial least squares, principal component regression, high dimensionality

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Definição dos objetivos

Uma das principais dificuldades encontradas por pesquisadores em análise de regressão é conseguir trabalhar com conjuntos de dados que possuem sérios problemas de multicolinearidade e/ou de alta dimensionalidade, em que o número de covariáveis é maior do que o número de observações. Estas situações podem causar resultados instáveis nos métodos de regressão por mínimos quadrados ordinários, além do aumento da variância dos coeficientes estimados.

Nesse sentido, vários autores têm trabalhado com o objetivo de desenvolver metodologias capazes de analisar dados com os problemas citados. Exemplos disso são as técnicas de regressão por componentes principais (Rawlings, Pentula e Dickey, 1998), por mínimos quadrados parciais (Wold *et al.*, 2001), por regressão ridge (Hoerl e Kennard, 1970a) e por LASSO (Tibshirani, 1996).

Basicamente, regressão por componentes principais (PCR) e mínimos quadrados parciais (PLS) resolvem o problema de multicolinearidade e de alta dimensionalidade criando "novas" variáveis, que são chamadas de fatores ou componentes. Os fatores, que conseguem explicar a(s) variável(is) resposta, capturam a variabilidade dos dados e são ortogonais entre si, resolvendo o possível problema de multicolinearidade. Obviamente, nem todos os fatores são utilizados, e à medida que um número menor destes componentes é selecionado, baseado em algum critério, maior é o percentual da variabilidade dos dados que é perdida.

Em relação a essas duas técnicas, a maior diferença entre elas é que o PLS consegue explicar uma mesma porcentagem da variabilidade dos dados com um número menor

de fatores, à medida que os mínimos quadrados parciais criam fatores decompondo o espaço das variáveis resposta e das covariáveis, enquanto que a regressão por componentes principais decompõe somente o espaço das covariáveis.

Por outro lado, regressão ridge e LASSO resolvem o problema de multicolinearidade acrescentando um pequeno vício ao estimador dos coeficientes de regressão encontrado por mínimos quadrados ordinários, afastando o sistema da singularidade. Tais metodologias, apesar de viciarem o estimador, possibilitam encontrar resultados com menores erros quadráticos médios, à medida que diminui consideravelmente a variância dos estimadores.

As principais diferenças entre a regressão ridge e o LASSO estão na penalização utilizada na expressão da soma de quadrados dos erros que deve ser minimizada para estimar os coeficientes de regressão; e no fato da regressão ridge levar as estimativas de alguns coeficientes à zero, enquanto que o LASSO faz com que as estimativas de alguns coeficientes sejam exatamente iguais a zero, tornando o segundo modelo mais interpretável, com menor número de covariáveis.

Assim sendo e de uma forma geral, neste trabalho pretendemos comparar essas quatro técnicas com relação ao poder de predição de cada uma delas para diferentes cenários de número de covariáveis, tamanho de amostra e quantidade e intensidade de coeficientes (efeitos) significativos, destacando as principais diferenças entre os métodos e possibilitando a criação de um guia para que o usuário possa escolher qual metodologia usar com base em algum conhecimento prévio que o mesmo possa ter.

1.2 Exemplos motivacionais

Esta seção tem o objetivo de apresentar exemplos motivacionais que requerem o uso das técnicas destacadas na definição dos objetivos.

Pesquisas em ciência e engenharia, muitas vezes, envolvem o uso de variáveis controláveis e/ou que são facilmente medidas, utilizadas para explicar ou prever o comportamento de outras variáveis (resposta). Quando essas variáveis controláveis estão em pequena quantidade, não são significativamente redundantes (colinearidade) e possuem uma relação bem compreendida com as variáveis resposta, a utilização de uma regressão linear múltipla pode ser uma boa maneira de transformar os dados em informações. No entanto, se pelo menos uma destas três condições se rompe, os modelos de regressão linear múltipla podem ser ineficientes ou inadequados. Em tais situações, outras técnicas devem

ser abordadas.

Em casos em que o pesquisador se depara com muitas variáveis e o objetivo é apenas construir um modelo preditivo, a regressão por mínimos quadrados parciais aparece como uma possível solução.

Suponha, por exemplo, que se tenha um processo químico qualquer cujo rendimento tem cinco componentes (variáveis) diferentes e seu objetivo é apenas prever as quantidades destes componentes baseadas em um espectro. Calibrado o instrumento a ser utilizado, você obtém 1000 frequências diferentes de 20 observações medidas das cinco componentes, que podem ser usadas para construir um modelo de predição linear. Para este caso, apresentado por Randall D. Tobias, do instituto SAS, uma regressão por mínimos quadrados parciais se torna adequada, possibilitando trabalhar com um exemplo típico de alta dimensionalidade ($p > n$) e de provável presença de problemas com multicolinearidade, em que as relações são mal entendidas.

No estudo em questão, o autor aplicou o método de regressão por mínimos quadrados parciais no *software* SAS e conseguiu obter um bom modelo preditivo com apenas 5 componentes PLS, explicando 85.69% da variabilidade das 1000 covariáveis e 99.54% da variabilidade das cinco variáveis resposta.

Um outro exemplo em que a utilização das técnicas foi de grande relevância, é no trabalho de Silva (2008). Em sua pesquisa são apresentadas alternativas para estimação na existência de multicolinearidade, em um conjunto de dados provenientes de um experimento conduzido pela empresa Agroindustrial Excelsior S.A. (Agrimex), localizada no Engenho Itapirema na cidade de Goiana - PE. Foram utilizadas 450 hastes de bambu, *Bambusa vulgaris*, que tiveram sua biomassa verde quantificada através do peso, em quilos. O interesse é predizer essa biomassa. Para isso, foram utilizadas quatro variáveis independentes medidas nas mesmas hastes. As variáveis independentes são: CAB^2H ($metros^3$), dada pelo produto entre a circunferência da base ao quadrado e a sua altura (medida muito utilizada como estimativa do volume em ciências florestais), a altura da haste (metros), a circunferência na base da haste (centímetros) e a circunferência a 1.30 metros do solo (centímetros).

Para este exemplo, Silva (2008), ao verificar a presença de problemas de multicolinearidade, utilizou diversos métodos alternativos ao método de regressão por mínimos quadrados ordinários para realizar as estimações. Dentre esses métodos estão a regressão ridge e a regressão por componentes principais. Para sua pesquisa, os métodos alter-

nativos de estimação conduziram a respostas semelhantes, mesmo possuindo estruturas diferentes. No entanto, a regressão por componentes principais apresentou os melhores resultados.

Vejamos agora o conjunto de dados do livro de Montgomery, Peck e Vining, de 2001, da Seção 9.2. Este conjunto de dados apresenta 16 observações em que o calor (em calorias por grama) de amostras de cimento é relacionado com quatro variáveis explicativas referentes a ingredientes usados na mistura do mesmo: (i) x_1, aluminato tricálcico, (ii) x_2, silicato tricálcico, (iii) x_3, aluminato-ferrita tetracálcico e (iv) x_4, silicato dicálcico. Para este exemplo, em uma básica análise exploratória, é possível destacar a seguinte matriz de correlações:

Tabela 1.1: Matriz de Correlações

	Calor	x_1	x_2	x_3	x_4
Calor	1	0.731	0.816	-0.535	-0.821
x_1		1	0.229	-0.824	-0.245
x_2			1	-0.139	-0.973
x_3				1	0.029
x_4					1

Analisando a Tabela 1.1 observamos a presença de grandes correlações, fornecendo um indicativo inicial da presença de multicolinearidade.

Assim sendo, ajustando um modelo de regressão por mínimos quadrados ordinários da forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

para $i = 1, ..., 16$, em que y_i denota o calor da i-ésima amostra de cimento e $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$, sendo independentes, obtemos as seguintes estimativas:

Tabela 1.2: Descrição das Estimativas

Efeito	Estimativa	Erro Padrão	valor-t	p-valor
Constante	62.405	70.071	0.89	0.40
x_1	1.551	0.745	2.08	0.07
x_2	0.510	0.724	0.70	0.50
x_3	0.102	0.755	0.13	0.90
x_4	-0.144	0.709	-0.20	0.84

Analisando a Tabela 1.2 observamos que apenas a covariável x_1 é marginalmente significativa. Entretanto, realizando uma análise de resíduos e de diagnóstico com o intuito de validar o modelo, é possível destacar a presença de duas observações influentes: 3 e 8.

Porém, quando retiramos a observação 3 da regressão, nenhum coeficiente fica significativo. Além disso, ocorre a mudança de sinal de dois coeficientes estimados. Já quando retiramos a observação 8 da regressão, o p-valor correspondente à estimativa do coeficiente da covariável x_1 fica menor, mas os demais coeficientes continuam não significativos e todos com sinal positivo.

Dessa forma, cabe ao pesquisador analisar e concluir se há de fato presença de multicolinearidade, à medida que os resultados são extremamente instáveis. Assim sendo, analisando a Tabela 1.3, logo a seguir, observamos a presença de grandes valores de VIF (Variance Inflation Factor) associados com todas as covariáveis consideradas. O VIF_k (VIF relacionado a k-ésima covariável) mede o quanto a variância do coeficiente da regressão padronizada é inflacionada por sua colinearidade e é um dos principais indicadores de problemas de colinearidade usados na literatura. Esta medida será detalhada de forma mais clara no próximo capítulo, mas, de uma forma geral, VIFs maiores do que 10 indicam sérios problemas de colinearidade.

Tabela 1.3: VIF_k das covariáveis

Variável	VIF
x_1	38.46
x_2	256.41
x_3	46.95
x_4	285.71

Logo, resta ao pesquisador a utilização de algum método paliativo, ou a utilização de alguma outra técnica de regressão que consiga trabalhar com este problema, como as técnicas discutidas na seção anterior.

Um outro exemplo é o conjunto de dados encontrado na biblioteca *pls* do *software* R, sendo carregado utilizando o comando *data(gasoline)*. Este banco de dados apresenta 60 observações e fornece o índice de octano de amostras de gasolina junto com outras 401 variáveis que medem o índice de espectros NIR (espectroscopia de infravermelho próximo) dessas amostras. O interesse é predizer o índice de octano (variável resposta) pelos índices de espectros NIR dessas 401 medidas de refletância.

Entretanto, para este exemplo, em que há alta dimensionalidade e o número de covariáveis é maior que o número de observações, nem um ajuste inicial com base nos métodos de regressão por mínimos quadrados ordinários se torna possível. Esses métodos são incapazes de trabalhar com situações nas quais o número de covariáveis é maior do que o número de observações.

Um último exemplo é um conjunto de dados sobre diabetes, originalmente usado por Efron, Hastie, Johnstone e Tibshirani (2004). Esse banco de dados apresenta informações sobre 422 pacientes diagnosticados com a doença. Para tais pacientes é registrado a sua idade, o seu sexo, o seu índice de massa corporal (IMC), a sua pressão arterial (PA) e seis medições de soro sanguíneo. A variável resposta é uma medida quantitativa da progressão da doença na data de referência. O objetivo é construir um modelo capaz de realizar previsões iniciais precisas da resposta para futuros pacientes.

Para este estudo, os pesquisadores comparam os resultados obtidos pelo LASSO com os resultados obtidos por uma regressão por mínimos quadrados ordinários. De uma forma geral, para um parâmetro de indexação (s) igual a 0.4, o LASSO selecionou as covariáveis sexo, índice de massa corporal, pressão arterial e duas medições de soro sanguíneo, que

obtiveram menores valores de erros padrões (calculados via técnicas de boostrap) do que os encontrados por mínimos quadrados ordinários, indicando que o LASSO prediz os coeficientes com maior acurácia.

Detalhes sobre alguns termos e expressões utilizados nesta seção, tais como o parâmetro de indexação s, logo atrás, serão detalhados nos próximos capítulos.

1.3 Revisão bibliográfica

Diversas pesquisas têm sido desenvolvidas com o intuito de comparar diferentes técnicas de regressão adequadas para situações em que há sérios problemas de multicolinearidade e/ou de alta dimensionalidade. Exemplos disso são Dormann *et al.* (2013) e Acharjee *et al.* (2013), que realizam revisões de diversas técnicas existentes, comparando-as por meio de simulações com respeito aos seus poderes preditivos.

Em Dormann *et al.* (2013) o foco é em casos com problemas de colinearidade, abordando principalmente estudos ecológicos. Os resultados obtidos indicaram uma preferência a utilização de métodos com penalização, tais como a regressão ridge e o LASSO, com relação aos métodos de variáveis latentes, tais como as regressões por componentes principais e por mínimos quadrados parciais. Além disso, é destacado que a escolha do método deve ser determinada por mais do que a sua capacidade de prever bem sob colinearidade. A possibilidade de obtenção de resultados interpretativos, por exemplo, é apresentada como de grande importância em estudos descritivos ecológicos, o que não se torna viável por meio da utilização do PCR e do PLS.

Já em Acharjee *et al.* (2013) o foco é em casos de alta dimensionalidade, em que o número de covariáveis é maior do que o número de observações. O estudo destaca como os métodos baseados em seleção de covariáveis, tais como o LASSO, selecionaram sete covariáveis simultaneamente, que seriam as principais dentro do seu estudo de simulação. Baseadas no erro quadrático médio de predição, as quatro técnicas que serão tratadas por este trabalho obtiveram resultados parecidos. Entretanto, o LASSO foi a que mais se destacou, com menor erro quadrático médio de predição. A pesquisa de Acharjee *et al.* (2013) destaca também a utilização de dados oriundos de um aplicativo da internet escrito em JAVA EE 6 e em Struts 2, que é executado por meio de um servidor específico.

Outros autores têm trabalhado em aperfeiçoar a escolha do parâmetro de penalização da regressão ridge, à medida que a abordagem mais amplamente adotada acabou por ser

apenas a inspeção visual do traço ridge (Hoerl e Kennard, 1970a), que traça as estimativas dessa técnica para cada valor da constante k de penalização. O valor escolhido seria aquele que representaria o mínimo valor a partir da qual as estimativas parecem se estabilizar.

Nesse contexto, Duzan e Shariff (2015) realizam uma revisão de diversos métodos elaborados de 1964 até 2014 para encontrar o parâmetro de penalização ideal. Já Wong e Chiu (2015) listam 26 métodos existentes em uma tabela, contribuindo com a criação de mais um, que minimiza o erro quadrático médio na regressão ridge por meio de uma abordagem iterativa.

Para este trabalho será utilizada a abordagem de Cule *et al.* (2012) para a escolha do parâmetro ridge. Este método é adequado para situações em que o número de covariáveis é maior do que o número de observações. Além disso, simulações realizadas pelos autores, com base neste método de escolha, apresentaram menores erros quadráticos médios do que o estimador de Hoerl, Kennard e Baldwin (1975) - que é amplamente utilizado na literatura e será detalhado nos próximos capítulos -, para casos em que há menos preditores do que observações e os erros de regressão não são tão pequenos.

Com relação ao LASSO, diversos estudos têm sido elaborados afim de encontrar a melhor forma de escolher o parâmetro de regularização associado ao método. Neste contexto, Tibshirani (1996) realiza diversas simulações para diferentes cenários de intensidade e quantidade de coeficientes (efeitos) significativos com o intuito de verificar quais das opções mais usuais têm os melhores resultados. De uma forma geral, os procedimentos baseados em validações cruzadas, que serão detalhados no próximo capítulo, foram os que produziram os melhores resultados, com menores erros quadráticos médios de predição. Logo, este trabalho utilizará desses procedimentos.

1.4 Apresentação do trabalho

Neste primeiro capítulo apresentamos a contextualização e a motivação deste trabalho, bem como seus objetivos principais, com uma revisão bibliográfica. No Capítulo 2 apresentamos as bases metodológicas da dissertação, necessárias para a compreensão das técnicas e dos procedimentos abordados. Já no terceiro capítulo são detalhadas as quatro técnicas citadas, que serão utilizadas e comparadas, destacando suas peculiaridades e principais características.

No quarto capítulo apresentamos as simulações realizadas, que são divididas em duas

partes. Na primeira é avaliado o poder de predição dos quatro métodos estudados para casos em que há problemas de multicolinearidade e o número de covariáveis é menor do que o de observações. Nesta abordagem, diferentes cenários de quantidade e intensidade de coeficientes (efeitos) significativos são considerados. Já na segunda parte é avaliado o poder de predição para casos em que há problemas de multicolinearidade e de alta dimensionalidade, com número de covariáveis maior do que o de observações. Nesta parte, diferentes tamanhos de amostra e número de covariáveis são fixados.

No quinto capítulo são destacadas aplicações em dados reais (não simulados). Já no sexto e último capítulo são apresentadas as considerações finais, em que os resultados obtidos ao longo do trabalho são discutidos.

1.5 Suporte computacional

As simulações e os gráficos apresentados foram construídos e elaborados no *software* R. A plataforma R é um ambiente de programação, análise de dados e gráficos, além de ser um *software* livre.

Capítulo 2

Metodologia

Este capítulo tem o objetivo de definir e apresentar as bases metodológicas que serão utilizadas para a construção deste trabalho.

O capítulo apresentará as principais definições necessárias para a compreensão e a utilização dos quatro métodos que serão comparados, tais como o que é centralização e escalonamento, amostra treinamento e amostra teste, e VIF (Variance Inflation Factor). Algumas medidas de acurácia/precisão para estimadores como erro quadrático médio e vício também serão definidas.

2.1 Centralização e Escalonamento

No decorrer deste trabalho utilizaremos variáveis e matrizes centradas e escalonadas. Dizemos que uma matriz $\mathbf{X}$ é centrada e escalonada se cada elemento dessa matriz é subtraído pela média da sua coluna correspondente e dividido pela raiz quadrada da soma de quadrados dos desvios com relação a média, ou seja, dividido por $\mathbf{S}_j = \sqrt{\sum_i (x_i - \overline{x})^2}$.

As principais motivações para o uso desse procedimento são:

- Reduzir o erro de arredondamento na inversão da matriz $\mathbf{X}^t\mathbf{X}$.
- Possibilitar a utilização das variáveis mais informativas para o modelo, dando a mesma importância para todas as variáveis antes da análise.
- Possibilitar a comparação direta dos coeficientes de regressão das diferentes variáveis.

Por exemplo: Suponha o modelo ajustado $\hat{y}$ = - 171 + 1.920 x_1 + 0.286 x_2 em que y, x_1 e x_2 representam a capacidade pulmonar em centilitros, a altura em centímetros e o peso

em quilogramas de indivíduos, respectivamente. Não faz sentido comparar os coeficientes 1.920 com 0.286, pois estão em diferentes escalas de medição. Agora, se centralizarmos e escalonarmos as covariáveis obteremos a seguinte equação estimada $\hat{y} = 193 + 12.90\ x_1 + 3.28\ x_2$. Dessa forma podemos comparar 12.90 e 3.28 para concluir que a diferença nas capacidades pulmonares são mais influenciadas pelas alturas do que pelos pesos.

2.2 Amostra treinamento/teste

Um procedimento bastante utilizado para medir a acurácia de modelos é dividir a amostra em amostra treinamento e em amostra teste.

Amostra treinamento é o nome dado ao conjunto de dados usado para a construção do modelo. Por construir o modelo entende-se estimar seus parâmetros e realizar uma seleção de covariáveis, de forma a se obter um modelo mais simples e interpretativo. De uma forma geral, a amostra treinamento possui cerca de 70% dos dados oriundos da amostra original, e as observações escolhidas para compor essa porcentagem são selecionadas por mecanismos aleatórios.

Amostra teste é o nome dado ao conjunto de dados usado para medir a acurácia do modelo criado, não tendo nenhum papel na construção do mesmo. Basicamente, após a construção do modelo, costuma-se prever a(s) variável(is) resposta da amostra teste utilizando o modelo que foi criado pela amostra treinamento. Os resultados dessas predições são comparados aos valores reais observados para essa(s) variável(is) utilizando alguma medida estatística para quantificar tais diferenças. Quanto menores forem os valores obtidos por essa medida, melhor será a acurácia do modelo construído. De uma forma geral, a amostra teste possui todas as observações que não são utilizadas para compor a amostra treinamento.

2.3 Validação cruzada

Validação cruzada é uma solução prática, confiável e altamente empregada para verificar o poder preditivo de um modelo. Basicamente, utiliza-se essa técnica para estimar o quão preciso algum modelo é na prática, ou seja, o seu desempenho para um novo conjunto de dados.

Diferentes métodos estatísticos utilizam validação cruzada como critério para a seleção

de uma quantidade de componentes ou para a escolha de um valor de uma constante de regularização. Exemplos disso são os métodos de regressão por componentes principais e por mínimos quadrados parciais, em que validações cruzadas são utilizadas para selecionar o número de componentes que deve ser retido, e os métodos de regressão ridge e LASSO, em que os valores de constantes são estabelecidos por meio dessa técnica. Essas quatro metodologias serão definidas no próximo capítulo.

O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutualmente exclusivos, utilizando alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento). O restante dos subconjuntos (dados de teste) são empregados na validação do modelo, medindo sua acurácia.

Um dos métodos de validação cruzada mais utilizados é o método k-fold. Basicamente, esse método funciona particionando o conjunto de dados em um número de grupos K, omitindo um grupo de cada vez. Os dados dos grupos não omitidos são utilizados para construir um modelo, e as respostas do grupo omitido são preditas, comparando os resultados encontrados com os valores reais observados, reservando os resíduos destas predições. Este processo é repetido até que cada grupo tenha sido omitido uma vez, e o total da soma dos quadrados destas diferenças seja calculado. Desta forma, temos então a estatística PRESS (Preditive REsidual Sum of Squares), que é a soma dos quadrados dos resíduos preditivos. Quanto menor for essa estatística, melhor será a acurácia (ou a performance) do modelo em questão.

Um caso específico do método k-fold é o método leave-one-out. Nesse caso, o número de grupos K é igual ao número de observações. Assim, cada grupo é constituído de apenas uma observação.

2.4 Multicolinearidade

Multicolinearidade consiste em um problema comum em regressões, no qual as variáveis independentes, ou covariáveis, possuem relações lineares exatas ou aproximadamente exatas. As consequências da multicolinearidade em uma regressão são a de erros-padrão elevados no caso de multicolinearidade moderada ou severa e até mesmo a impossibilidade de se obter estimativas se a multicolinearidade for perfeita.

Alguns sintomas que podem indicar um problema de colinearidade são:

- grandes mudanças em $\hat{\beta}_k$ quando adicionamos ou excluímos covariáveis ou observações;
- β_k não significante para um X_i teoricamente importante;
- $\hat{\beta}_k$ com sinal oposto ao esperado;
- correlações grandes na matriz de correlações de $\mathbf{X}$;
- valores muito grandes de $s(\hat{\beta}_k)$;
- β_k's não significantes mesmo quando o teste F da regressão é significante.

2.4.1 Fator de Inflação da Variância (VIF)

Os elementos da diagonal principal de $(\mathbf{X}'\mathbf{X})^{-1}$ são úteis para detectar multicolinearidade. O k-ésimo elemento da diagonal principal de $(\mathbf{X}'\mathbf{X})^{-1}$, C_{kk}, pode ser escrito como:

$$C_{kk} = (1 - R_k^2)^{-1}, \quad k = 1, ..., p-1,$$

em que R_k^2 é o R^2 da regressão de X_k sobre as outras covariáveis. Se X_k é aproximadamente ortogonal às outras covariáveis, então R_k^2 é pequeno e C_{kk} é próximo de 1. C_{kk} é chamado de fator de inflação da variância e uma outra notação usada é VIF_k.

Mas por que a expressão fator de inflação da variância?

O modelo de regressão padronizada é um modelo de regressão no qual todas as variáveis (X e Y) são padronizadas em escores z, com média 0 e desvio padrão 1, e divididas por $(n-1)^{\frac{1}{2}}$. Em um modelo desses, as equações normais $(\mathbf{X}'\mathbf{X})^{-1}\hat{\beta} = (\mathbf{X}'\mathbf{Y})$ reduzem-se a:

$$r_{\mathbf{XX}}\hat{\beta}^* = r_{\mathbf{XY}},$$

em que $r_{\mathbf{XX}}$ é a matriz de correlações de $\mathbf{X}$, $r_{\mathbf{XY}}$ são as correlações de $\mathbf{Y}$ com $\mathbf{X}$ e $\hat{\beta}^*$ é o vetor dos coeficientes de regressão padronizados.

Pode ser mostrado que VIF_k é o k-ésimo elemento da diagonal de ${r_{\mathbf{XX}}}^{-1}$, de forma que:

$$var(\hat{\beta}_k^*) = (\sigma^*)^2 \;\; VIF_k = \frac{(\sigma^*)^2}{(1 - R_k^2)},$$

em que $(\sigma^*)^2$ é a variância do erro do modelo padronizado. Assim, verifica-se que VIF_k mede o quanto a variância do coeficiente da regressão padronizada, $\hat{\beta}_k^*$, é inflacionada por sua colinearidade.

De uma forma geral, considera-se que VIF's maiores do que 10 indicam sérios problemas de colinearidade.

2.5 Propriedades frequentistas dos estimadores

Visando a continuidade da pesquisa, esta seção aborda algumas propriedades frequentistas dos estimadores que podem ser utilizadas para verificar a qualidade e a funcionalidade dos mesmos.

2.5.1 Erro Quadrático Médio, Variância e Vício

Para medir o desempenho de um estimador, alguns conceitos serão definidos e algumas propriedades serão apresentadas. Uma dessas propriedades é o erro quadrático médio (EQM), que é definido como a esperança da diferença ao quadrado entre o parâmetro e o seu estimador, podendo ser representado por:

$$EQM[\theta, \hat{\theta}] = E[(\hat{\theta} - \theta)^2],$$

na qual θ é o parâmetro que deve ser estimado e $\hat{\theta}$ é o seu estimador.

Outra propriedade importante é a variância de um estimador, ou seja, a medida que calcula o quanto esse estimador é preciso. A variância de um estimador pode ser representada por $V[\hat{\theta}]$, e, calculada, da seguinte forma:

$$V[\hat{\theta}] = E[\hat{\theta}^2] - E^2[\hat{\theta}].$$

Além disso, existe o vício, que pode ser representado por $B[\theta, \hat{\theta}]$. O vício de um estimador é a média da diferença do estimador para o seu parâmetro, ou seja, é calculado pela seguinte expressão:

$$B[\theta, \hat{\theta}] = E[\hat{\theta} - \theta].$$

Um estimador é visto como bom e preciso quando as medidas de desempenho, caracterizadas pelos valores calculados por essas propriedades, possuem resultados absolutos pequenos. O vício, por exemplo, caracteriza um estimador em viciado ou não viciado. Um estimador é dito não viciado quando possui vício igual a 0, sendo a média da diferença do estimador para o seu parâmetro, nula.

2.5.2 Relação Funcional entre as Medidas

Existe uma relação funcional entre o EQM, a variância e o vício de um estimador, e que será demonstrada nesta subseção.

Retomando, vimos que podemos representar o EQM como:

$$EQM[\theta, \hat{\theta}] = E[(\hat{\theta} - \theta)^2]$$

$$= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2]$$

$$= E[(\hat{\theta} - E[\hat{\theta}])^2 + 2\,(\hat{\theta} - E[\hat{\theta}])\,(E[\hat{\theta}] - \theta) + (E[\hat{\theta}] - \theta)^2]$$

$$= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] - \theta)^2$$

$$= V[\hat{\theta}] + B^2[\theta, \hat{\theta}].$$

Logo, a relação funcional entre o EQM, a variância e o vício de um estimador é:

$$EQM[\theta, \hat{\theta}] = V[\hat{\theta}] + B^2[\theta, \hat{\theta}].$$

Essa relação e várias explicações com exemplos ilustrativos dessas propriedades frequentistas dos estimadores podem ser vistas em Bolfarine & Sandoval (2001) e Bussab & Morettin (1987).

Capítulo 3

Técnicas estatísticas

Este capítulo tem o objetivo de apresentar as quatro técnicas abordadas na introdução, adequadas para situações em que há sérios problemas de multicolinearidade e/ou de alta dimensionalidade.

3.1 Regressão por Componentes Principais (PCR)

Esta seção inicia as apresentações das técnicas que serão comparadas. Assim sendo, a primeira técnica discutida é a de regressão por componentes principais (PCR).

Regressão por componentes principais aborda o problema de colinearidade diminuindo as dimensões do espaço de $\mathbf{X}$ que estão causando o problema. Isto é semelhante, em conceito, a retirar uma variável independente do modelo quando esta apresenta dispersão insuficiente para contribuir com informações significativas sobre $\mathbf{Y}$. No entanto, em PCR, a dimensão desconsiderada é definida por uma combinação linear das covariáveis, ao invés de apenas uma variável independente.

Na construção do modelo de regressão por componentes principais devemos considerar a matriz $\mathbf{X}$ centrada e escalonada. Assim, utilizando uma decomposição em valores singulares (SVD), devemos decompor esta matriz na forma $\mathbf{X} = \mathbf{U}\mathbf{L}^{1/2}\mathbf{V}^t$, em que $\mathbf{U}$ é uma matriz n x p e $\mathbf{V}$ é uma matriz p x p contendo os vetores singulares à esquerda e à direita de $\mathbf{X}$, respectivamente, tal que $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$. $\mathbf{L}^{1/2}$ é uma matriz diagonal de valores singulares. Nessa abordagem, os valores singulares e seus respectivos autovetores são ordenados tal que $\lambda_1^{1/2} > \lambda_2^{1/2} > ... > \lambda_p^{1/2}$ (elementos da diagonal de $\mathbf{L}^{1/2}$).

Os componentes principais de $\mathbf{X}$ são definidos como funções lineares das covariáveis especificados pelos coeficientes dos vetores colunas de $\mathbf{V}$. O primeiro autovetor de $\mathbf{V}$ (pri-

meira coluna) define o primeiro componente, o segundo autovetor de $\mathbf{V}$ define o segundo componente, e assim por diante. Cada componente principal é uma função linear de todas as variáveis independentes. Os componentes principais são dados também pelas colunas de $\mathbf{U}$ multiplicado pelos correspondentes $\lambda_j^{1/2}$. Assim,

$$\mathbf{W} = \mathbf{XV} = \mathbf{UL}^{1/2}$$

é a matriz de componentes principais. Cada coluna em $\mathbf{W}$ fornece os valores para cada uma das n observações em cada um dos componentes principais.

A soma de quadrados dos componentes principais $\mathbf{W}$ é a matriz diagonal dos autovalores,

$$\mathbf{W}^t\mathbf{W} = (\mathbf{UL}^{1/2})^t(\mathbf{UL}^{1/2}) = \mathbf{L}^{1/2}\mathbf{U}^t\mathbf{UL}^{1/2} = \mathbf{L},$$

em que $\mathbf{L} = \text{Diag}(\lambda_1,\ \lambda_2,\ ...,\ \lambda_p)$. Assim, a soma de quadrados de cada componente principal é igual ao seu correspondente autovalor λ_j. O primeiro componente principal tem a maior soma de quadrados, λ_1. Os componentes principais correspondentes aos menores autovalores são as dimensões do espaço de $\mathbf{X}$ que tem as menores dispersões. Estas dimensões do espaço de $\mathbf{X}$ com dispersões limitadas são responsáveis pelo problema de colinearidade quando o mesmo existe.

3.1.1 O modelo linear

O modelo linear $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, em que $\mathbf{X}$ está centrada e escalonada e $\mathbf{1}$ representa um vetor unitário, ou seja, constituído apenas de uns, pode ser reescrito em termos dos componentes principais $\mathbf{W}$ como:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

em que os erros ϵ_i são independentes e possuem variância constante σ^2.

Para encontrar esta reformulação do modelo, basta utilizar o fato de que $\mathbf{VV}^t = \mathbf{I}$, transformando $\mathbf{X}\boldsymbol{\beta}$ em $\mathbf{W}\boldsymbol{\gamma}$ da seguinte forma:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{XVV}^t\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}.$$

Note que $\boldsymbol{\gamma} = \mathbf{V}^t\boldsymbol{\beta}$ é o vetor de coeficientes de regressão para os componentes principais. Assim, para retornar aos coeficientes de regressão de $\mathbf{X}$, basta fazer $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\gamma}$.

Assim sendo, o estimador de mínimos quadrados ordinários de $\boldsymbol{\gamma}$ utilizando todos os componentes principais como variáveis independentes é dado por:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{W}^t\mathbf{W})^{-1}\mathbf{W}^t\mathbf{Y} = \mathbf{L}^{-1}\mathbf{W}^t\mathbf{Y}.$$

Os coeficientes de regressão para os componentes principais podem ser computados individualmente já que os componentes principais são ortogonais: $\mathbf{W}^t\mathbf{W}$ é uma matriz diagonal. Do mesmo modo, a matriz de variâncias-covariâncias de $\hat{\boldsymbol{\gamma}}$ é a matriz diagonal $\mathbf{Var}(\hat{\boldsymbol{\gamma}}) = \mathbf{L}^{-1}\sigma^2$. Isto é, a variância de $\hat{\gamma}_j$ é $\sigma^2(\hat{\gamma}_j) = \frac{\sigma^2}{\lambda_j}$ e todas as covariâncias são iguais a zero.

Assim, apesar de não ser uma prática usual, se todos os componentes principais são usados, os resultados são equivalentes aos obtidos por mínimos quadrados ordinários. Neste caso, a estimativa de $\boldsymbol{\beta}$ é obtida de $\hat{\boldsymbol{\gamma}}$ como $\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$ e a equação de regressão pode ser reescrita como $\hat{\mathbf{Y}} = \mathbf{1}\overline{Y} + \mathbf{W}\hat{\boldsymbol{\gamma}}$ ou $\hat{\mathbf{Y}} = \mathbf{1}\overline{Y} + \mathbf{X}\hat{\boldsymbol{\beta}}$.

3.1.2 Seleção do número de componentes

Alguns dos objetivos por trás da regressão por componentes principais são diminuir a dimensionalidade de $\mathbf{X}$ e solucionar o problema de colinearidade, transformando as dimensões que estão causando este problema, ou seja, aquelas com pequenos valores de λ_j. Assim, alguma estratégia para a seleção dos componentes que devem ser utilizados na regressão precisa ser utilizada.

Uma estratégia para a escolha de M, o número de componentes retidos, é simplesmente deletar todos os componentes que possuem variância menor que l^*, em que l^* é algum nível de corte. A escolha de l^* é bastante arbitrária, mas quando se lida com matrizes de correlação, em que o valor médio dos autovalores é 1, um valor de l^* dentro do intervalo de 0,01 a 0,1 parece ser bem útil na prática (Jolliffe, 2002).

Hill *et al.* (1977) realiza uma abrangente discussão de maneiras mais sofisticadas de escolher quais e quantos componentes devem ser retidos. Entretanto, como o foco deste trabalho será baseado em obter boas predições para $\mathbf{Y}$, consideraremos a estratégia de Mertens *et al.* (1995), que utiliza uma versão da estatística PRESS, que representa a soma dos quadrados dos resíduos preditivos, como um critério para decidir quantos componentes

principais devem ser retidos para a regressão. O seu critério calcula

$$\sum_{i=1}^{n}(y_i - \hat{y}_{M(i)})^2,$$

em que $\hat{y}_{M(i)}$ é a predição de y_i obtida pela regressão por componentes principais baseada em um subconjunto M de componentes e usando a matriz de dados $\mathbf{X}_{(i)}$, na qual $\mathbf{X}$ está com a sua linha i deletada. O subconjunto de componentes principais que obtiver o menor valor para esta estatística é, assim, retido para a regressão. É possível observar que tal procedimento nada mais é do que uma validação cruzada.

Assim sendo, selecionado o número de componentes, podemos calcular o estimador dos coeficientes para $\mathbf{X}$ da seguinte forma:

$$\boldsymbol{\beta}^{+}_{(m)} = \mathbf{V}_{(m)}\hat{\boldsymbol{\gamma}}_{(m)}$$

em que $\boldsymbol{\beta}^{+}_{(m)}$ representa as estimativas de $\boldsymbol{\beta}$ da regressão de $\mathbf{Y}$ pelos M componentes retidos e $\mathbf{V}_{(m)}$ e $\hat{\boldsymbol{\gamma}}_{(m)}$ representam as partições de $\mathbf{V}$ e $\hat{\boldsymbol{\gamma}}$, respectivamente, que se relacionam com as dimensões que compõem os M componentes principais selecionados.

A equação de regressão para os M componentes pode ser escrita como $\hat{\mathbf{Y}}_{(m)} = \mathbf{1}\overline{Y} + \mathbf{W}_{(m)}\hat{\boldsymbol{\gamma}}_{(m)}$ ou $\hat{\mathbf{Y}}_{(m)} = \mathbf{1}\overline{Y} + \mathbf{X}\boldsymbol{\beta}^{+}_{(m)}$, em que $\mathbf{W}_{(m)}$ é a matriz dos componentes principais retidos.

3.2 Regressão por mínimos quadrados parciais (PLS)

Esta seção apresenta o segundo método que será utilizado. O método de regressão por mínimos quadrados parciais (PLS).

Modelos de regressão por mínimos quadrados parciais (Wold *et al.*, 2001) são alternativas para modelar as relações existentes entre conjuntos de variáveis observadas por meio de variáveis latentes, quando os tamanhos amostrais são pequenos com relação ao número de variáveis independentes ou ainda em situações em que ocorre alto grau de correlação entre tais covariáveis (multicolinearidade). Nos métodos de regressão por mínimos quadrados ordinários, tais situações podem gerar resultados instáveis, além do aumento da variância dos coeficientes estimados.

3.2.1 Descrição do método PLS

Assim como em regressão linear múltipla multivariada, a principal finalidade da regressão por mínimos quadrados parciais é construir um modelo linear $\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{E}$, em que $\mathbf{Y}$ é uma matriz n x m de variáveis resposta, $\mathbf{X}$ é uma matriz n x p de variáveis preditoras, $\boldsymbol{\beta}$ é uma matriz p x m dos coeficientes de regressão, e $\mathbf{E}$ é a matriz de ruídos para o modelo, que tem a mesma dimensão de $\mathbf{Y}$. Os erros no modelo de regressão PLS têm as mesmas suposições que na regressão linear múltipla, exceto pela distribuição. PLS é uma abordagem livre de distribuição. A principal consequência disso é que os estimadores dos coeficientes da regressão não possuem distribuições conhecidas. Logo, são necessárias técnicas de reamostragem como boostrap para verificar a significância dos coeficientes.

O método de regressão PLS extrai um pequeno número de "novas" variáveis, que são chamadas de fatores ou componentes e são denotadas por $\boldsymbol{t_i}$ (i = 1, ..., a). Tais fatores são preditores de $\mathbf{Y}$ e também descrevem $\mathbf{X}$, isto é, tanto $\mathbf{X}$ como $\mathbf{Y}$ são, pelo menos em parte, modelados pelas mesmas variáveis latentes.

A ideia do método é extrair fatores que consigam capturar a variabilidade das covariáveis sendo, também, bons preditores das variáveis resposta. Isto pode ser conseguido modificando as variáveis latentes de forma que as covariâncias entre os componentes de $\mathbf{X}$, $\boldsymbol{t_i}$, e $\mathbf{Y}$ sejam maximizadas.

O número de componentes extraídos de $\mathbf{X}$ é menor que o número de covariáveis, ou seja, a é menor que p, e os mesmos são ortogonais. Estes componentes são obtidos como combinações lineares das variáveis originais $\boldsymbol{x_k}$, com os coeficientes, "pesos", $\boldsymbol{w_i}$ (i = 1, ..., a), dados por:

$$\mathbf{T} = \mathbf{XW}, \tag{3.1}$$

em que $\mathbf{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_a)$ é uma matriz n x a de fatores e $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_a)$ é uma matriz p x a de pesos. Assim, as matrizes $\mathbf{X}$ e $\mathbf{Y}$ são decompostas, como em uma análise fatorial, ou seja,

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{F}$$

$$\mathbf{Y} = \mathbf{UC}^t + \mathbf{G},$$

sendo que $\mathbf{T}$ e $\mathbf{U}$ são matrizes n x a de fatores de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente; $\mathbf{P}^t$ é a matriz de cargas de $\mathbf{X}$ de dimensão a x p; $\mathbf{C}^t$ é a matriz de cargas de $\mathbf{Y}$ de dimensão a x m; e $\mathbf{F}$ e $\mathbf{G}$ são matrizes de erros.

Como já foi comentado, na decomposição de $\mathbf{X}$ as componentes, $\boldsymbol{t_i}$, são obtidas de maneira que as covariâncias entre elas e as variáveis resposta do modelo, da matriz $\mathbf{Y}$, sejam maximizadas.

Com a dimensão de $\mathbf{X}$ reduzida em a componentes $\boldsymbol{t_i}$ $(a < p)$ pode-se efetuar a regressão de $\mathbf{Y}$ sobre $\mathbf{T}$ na forma:

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^t + \mathbf{H}. \tag{3.2}$$

Para conseguir os "coeficientes da regressão PLS", basta substituir a igualdade em 3.1, na equação 3.2, e obter:

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^t + \mathbf{H} = \mathbf{X}\mathbf{W}\mathbf{C}^t + \mathbf{H} = \mathbf{X}\mathbf{B} + \mathbf{H}.$$

Assim, os coeficientes da regressão PLS podem ser escritos como $\mathbf{B} = \mathbf{W}\mathbf{C}^t$.

O estimador para $\mathbf{C}$ é obtido por mínimos quadrados, e é dado por:

$$\hat{\boldsymbol{C}}^t = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{Y}.$$

Consequentemente,

$$\hat{\mathbf{B}} = \boldsymbol{W}\hat{\boldsymbol{C}}^t = \begin{bmatrix} \hat{b}_{0,1} & \dots & \hat{b}_{0,m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \hat{b}_{p,1} & \dots & \hat{b}_{p,m} \end{bmatrix}.$$

A j-ésima coluna da matriz $\hat{\mathbf{B}}$ corresponde aos coeficientes estimados para o modelo referente à variável resposta $\mathbf{Y}_j$, j = 1, 2, ..., m.

3.2.2 O Algoritmo NIPALS

Existem diversos algoritmos que são utilizados para o cálculo das componentes da regressão PLS. Um dos mais conhecidos é o algoritmo Nonlinear Iterative Partial Least Squares (NIPALS), que foi desenvolvido originalmente por Wold (1966). O algoritmo é

descrito abaixo, e trabalha com as matrizes de dados originais $\mathbf{X}$ e $\mathbf{Y}$ padronizadas, ou seja, escalonadas e centradas em zero.

1. Faça $\mathbf{u}$ igual a uma das colunas de $\mathbf{Y}$;
2. Determine uma coluna dos pesos de $\mathbf{W}$, utilizando $\mathbf{w} = \mathbf{X}^t\mathbf{u}/\mathbf{u}^t\mathbf{u}$;
3. Calcule uma coluna dos $\mathbf{T}$, por meio de $\mathbf{t} = \mathbf{X}\mathbf{w}$;
4. Determine os pesos de $\mathbf{Y}$, $\mathbf{c}$, usando $\mathbf{c} = \mathbf{Y}^t\mathbf{t}/\mathbf{t}^t\mathbf{t}$;
5. Faça a atualização dos fatores de $\mathbf{Y}$, $\mathbf{u}$, através de $\mathbf{u} = \mathbf{Y}\mathbf{c}/\mathbf{c}^t\mathbf{c}$;
6. Teste a convergência de $\mathbf{t}$, isto é, $||\boldsymbol{t}_{velho} - \boldsymbol{t}_{novo}||/||\boldsymbol{t}_{novo}|| \leq \xi$, em que ξ é uma constante predeterminada e pequena, 10^{-6} ou 10^{-8}, por exemplo. Se não houver convergência retorne ao passo 2, caso contrário siga para o passo 7;
7. Faça

$$\mathbf{p} = \mathbf{X}^t\mathbf{t}/\mathbf{t}^t\mathbf{t}$$

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^t$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{c}^t;$$

8. Continue com o próximo componente, retornando ao passo 1, até que o critério utilizado para a escolha do número de componentes (uma validação cruzada, por exemplo) indique que a quantidade está adequada.

3.2.3 Diferenças, vantagens e desvantagens do PLS

Esta seção tem o intuito de resumir de forma clara as principais diferenças entre o método de mínimos quadrados ordinários e o método de mínimos quadrados parciais (PLS).

O método de mínimos quadrados ordinários, diferentemente do PLS, pode apresentar resultados instáveis para tamanhos de amostra pequenos em relação ao número de variáveis independentes. Além disso, o alto grau de correlação entre as covariáveis (multicolinearidade) pode aumentar a variância dos coeficientes estimados no método de mínimos quadrados ordinários, o que não ocorre no PLS, à medida que os componentes criados, utilizados na regressão, são ortogonais.

A seguir, são resumidas as principais vantagens e desvantagens da utilização do PLS com relação ao método de mínimos quadrados ordinários.

Vantagens:

- O método é hábil para modelar regressões com múltiplas variáveis resposta;
- Não é afetado por multicolinearidade;
- Produz fatores que apresentam elevadas covariâncias com as variáveis resposta, ou seja, fatores com alto poder de predição.

Desvantagens:

- Dificuldade na interpretação das cargas fatoriais, servindo, geralmente, apenas para a predição de novas observações;
- Os estimadores dos coeficientes de regressão não possuem distribuições conhecidas e, com isso, o teste de significância dos mesmos só pode ser realizado via métodos de reamostragem. Exemplo: boostrap;
- Falta de estatísticas de teste para o modelo.

3.3 Regressão Ridge (RR)

O modelo de regressão ridge (Hoerl e Kennard, 1970) objetiva eliminar a multicolinearidade das variáveis explicativas adicionando uma pequena quantidade positiva no estimador $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$, ou seja, viciando o mesmo na forma:

$$\boldsymbol{\beta}^* = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{Y}. \tag{3.3}$$

Assim sendo, obter 3.3 é o mesmo que minimizar a soma de quadrados penalizada a seguir:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + k\sum_{j=1}^{p}\beta_j^2,$$

que é equivalente a minimizar $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ sujeito a algum $c > 0$, tal que $\sum_{j=1}^{p}\beta_j^2 < c$, isto é, limitando a soma de quadrados dos coeficientes e fazendo com o que os mesmos tendam a 0 para maiores valores de k.

A ideia por trás do método é obter um estimador que tenha um pequeno vício, mas que possua uma variância menor que a dos mínimos quadrados, à medida que o teorema de Gauss Markov garante variância mínima somente dentre os estimadores não viciados, mas não garante que esta seja a menor possível em qualquer situação.

O problema causado pela multicolinearidade é resolvido afastando a singularidade da matriz $\mathbf{X}^t\mathbf{X}$ por meio do acréscimo da constante k, possibilitando o cálculo da inversa dentro da fórmula do estimador $\hat{\boldsymbol{\beta}}$.

A relação entre o estimador ridge com o estimador de mínimos quadrados ordinários pode ser ilustrada. Sabemos que $\boldsymbol{\beta}^* = (\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{Y}$, assim sendo, denotando $(\mathbf{X}^t\mathbf{X}+k\mathbf{I})^{-1}$ por $\mathbf{F}$, teremos que $\boldsymbol{\beta}^* = \mathbf{F}\mathbf{X}^t\mathbf{Y}$. Agora, utilizando a igualdade dos sistemas de equações normais $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$, teremos que:

$$\boldsymbol{\beta}^* = \mathbf{F}\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} =$$

$$= \left[(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X}) + k(\mathbf{X}^t\mathbf{X})^{-1}\right]^{-1}\hat{\boldsymbol{\beta}} \;=\; \left[\mathbf{I} + k(\mathbf{X}^t\mathbf{X})^{-1}\right]^{-1}\hat{\boldsymbol{\beta}}.$$

Dessa forma, denotando $\left[\mathbf{I} + k(\mathbf{X}^t\mathbf{X})^{-1}\right]^{-1}$ por $\mathbf{Z}$, teremos que $\boldsymbol{\beta}^* = \mathbf{Z}\hat{\boldsymbol{\beta}}$. Logo, o valor esperado do estimador é:

$$E(\boldsymbol{\beta}^*) = E(\mathbf{Z}\hat{\boldsymbol{\beta}}) = \mathbf{Z}E(\hat{\boldsymbol{\beta}}) = \mathbf{Z}\boldsymbol{\beta}.$$

Assim sendo, teremos que $\boldsymbol{\beta}^*$ é um estimador viciado se $\mathbf{Z} \neq \mathbf{I}$, já que $\mathbf{Z}$ é uma matriz que depende de k. Se $\mathbf{Z} = \mathbf{I}$, teremos que $k = 0$, o que nos fornece o estimador não viciado de mínimos quadrados.

3.3.1 Propriedades

Apresentamos a seguir algumas propriedades importantes de $\boldsymbol{\beta}^*$, $\mathbf{F}$ e $\mathbf{Z}$:

1. Sejam $\xi_i(\mathbf{F})$ e $\xi_i(\mathbf{Z})$ os autovalores de $\mathbf{F}$ e $\mathbf{Z}$, respectivamente. Então,

$$\xi_i(\mathbf{F}) = \frac{1}{\lambda_i + k} \;\; e \;\; \xi_i(\mathbf{Z}) = \frac{\lambda_i}{\lambda_i + k},$$

em que λ_i para $i = 1, 2, ..., p$ são os autovalores de $\mathbf{X}^t\mathbf{X}$.

2. $\mathbf{Z}$ pode ser escrito na forma $\mathbf{Z} = \mathbf{I}$ - $k(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1} = \mathbf{I}$ - $k\mathbf{F}$.

3. Para $k \neq 0$, $\boldsymbol{\beta}^*$ tem norma menor que $\hat{\boldsymbol{\beta}}$, isto é, $\boldsymbol{\beta}^{*t}\boldsymbol{\beta}^* < \hat{\boldsymbol{\beta}}^t\hat{\boldsymbol{\beta}}$.

3.3.2 Erro quadrático médio dos estimadores ridge

Nesta seção apresentamos a fórmula do erro quadrático médio do estimador ridge para uma dada constante k, que é calculada com base na distância entre $\boldsymbol{\beta}^*(k)$ e $\boldsymbol{\beta}$. Dessa forma, se denotarmos o erro quadrático médio deste estimador por $EQM(k)$, teremos que $EQM(k) = E((\boldsymbol{\beta}^*(k) - \boldsymbol{\beta})^t(\boldsymbol{\beta}^*(k) - \boldsymbol{\beta}))$. Assim, somando e subtraindo o termo $E(-2\hat{\boldsymbol{\beta}}^t\mathbf{Z}^t\mathbf{Z}\boldsymbol{\beta} + 2\boldsymbol{\beta}^t\mathbf{Z}^t\mathbf{Z}\boldsymbol{\beta} - 2\boldsymbol{\beta}^t\mathbf{Z}^t\boldsymbol{\beta})$ na expressão anterior, teremos que:

$$EQM(k) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\mathbf{Z}^t\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) + (\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta})^t(\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta}). \tag{3.4}$$

Em 3.4, o segundo termo é a distância ao quadrado de $\mathbf{Z}\boldsymbol{\beta}$ a $\boldsymbol{\beta}$. Assim, pode ser considerado como o quadrado do vício. O primeiro termo, veremos mais adiante que é a soma das variâncias (variância total) dos estimadores dos parâmetros. Desenvolvendo cada um dos termos, teremos que:

$$EQM(k) = E(tr((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\mathbf{Z}^t\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))) + \boldsymbol{\beta}^t(\mathbf{Z} - \mathbf{I})^t(\mathbf{Z} - \mathbf{I})\boldsymbol{\beta} =$$

$$= E(tr(\mathbf{Z}^t\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t)) + \|(\mathbf{Z} - \mathbf{I})\boldsymbol{\beta}\|^2 .$$

Usando a segunda propriedade temos que $\mathbf{Z}$ - $\mathbf{I} = -k\mathbf{F}$ em que $\mathbf{F} = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}$. Logo, teremos a seguinte igualdade:

$$EQM(k) = tr(\mathbf{Z}^t\mathbf{Z}E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t)) + \|(-k)\mathbf{F}\boldsymbol{\beta}\|^2 =$$

$$= tr(\mathbf{Z}^t\mathbf{Z}Var(\hat{\boldsymbol{\beta}})) + k^2\|\mathbf{F}\boldsymbol{\beta}\|^2 =$$

$$= \sigma^2 tr(\mathbf{Z}^t\mathbf{Z}(\mathbf{X}^t\mathbf{X})^{-1}) + k^2\left\|(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\boldsymbol{\beta}\right\|^2 .$$

Como $\mathbf{Z} = \mathbf{F}\mathbf{X}^t\mathbf{X}$ implica que $\mathbf{Z}(\mathbf{X}^t\mathbf{X})^{-1} = \mathbf{F}$, tem-se que:

$$EQM(k) = \sigma^2 tr(\mathbf{Z}^t\mathbf{F}) + k^2 tr[\boldsymbol{\beta}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta}] =$$

$$= \sigma^2 tr(\mathbf{Z}^t\mathbf{F}) + k^2 tr[\boldsymbol{\beta}^t\boldsymbol{\beta}(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-2}].$$

Agora, fazendo a transformação $\boldsymbol{\alpha} = \mathbf{V}\boldsymbol{\beta}$ (SVD), temos que $\boldsymbol{\beta}^t\boldsymbol{\beta} = \boldsymbol{\alpha}^t\mathbf{V}\mathbf{V}^t\boldsymbol{\alpha} = \boldsymbol{\alpha}^t\boldsymbol{\alpha}$. Assim:

$$EQM(k) = \sigma^2 tr(\mathbf{Z}^t\mathbf{F}) + k^2 tr[\boldsymbol{\alpha}^t\boldsymbol{\alpha}(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-2}].$$

Utilizando a primeira propriedade, temos que $\xi_i(\mathbf{F}) = \frac{1}{\lambda_i + k}$ e $\xi_i(\mathbf{Z}) = \frac{\lambda_i}{\lambda_i + k}$. Como $\mathbf{Z}$ e $\mathbf{F}$ são matrizes diagonais teremos que $\xi_i(\mathbf{ZF}) = \frac{\lambda_i}{(\lambda_i + k)^2}$. Logo, segue que:

$$EQM(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{{\alpha_i}^2}{(\lambda_i + k)^2} = \gamma_1(k) + \gamma_2(k)$$

em que $\gamma_1(k) + \gamma_2(k)$ são, respectivamente, a variância total e o vício-quadrado do estimador ridge.

A figura 3.1, retirada do artigo de Hoerl e Kennard (1970a), mostra o esboço do comportamento das funções $\gamma_1(k) + \gamma_2(k)$ e a soma de ambas.

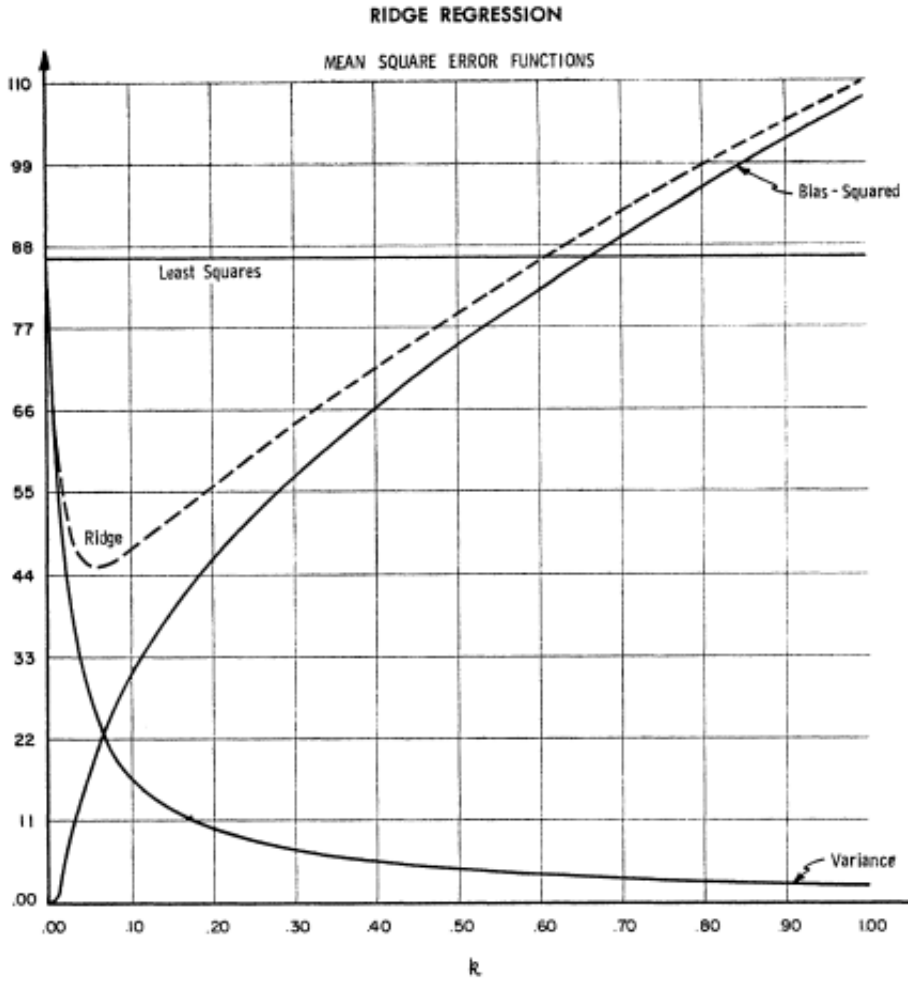

Figura 3.1: Variância, vício-quadrado e a soma de ambos, EQM(k), como função de k (Hoerl e Kennard, 1970a)

Analisando a Figura 3.1 verificamos que quando $k = 0$ o estimador ridge é igual ao estimador de mínimos quadrados ordinários, com vício-quadrado igual a 0 e variância igual a $\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$. Além disso, a medida que cresce o valor de k aumenta o valor do vício e diminui o valor da variância, nos levando a estimadores viciados, mas com variância menor que a dos mínimos quadrados. Como indicado pelo gráfico, a soma do vício-quadrado e da variância resulta na soma dos erros quadráticos médio, e que a medida que k cresce o EQM(k) diminui para um mínimo global, menor que o EQM dos estimadores de mínimos quadrados, voltando novamente a crescer quando k é muito grande. Este mínimo global é o que justifica a utilização do método, pois possibilita encontrar estimadores com erro quadrático médio menor que os de mínimos quadrados ordinários.

3.3.3 Estimação de k

Para a estimação da constante k apresentamos a proposta de Cule *et al.* (2012). Esta proposta é adequada para situações de alta dimensionalidade, em que o número de covariáveis é maior do que o número de observações ($p > n$).

Com as colunas de $\mathbf{X}$ centradas e escalonadas, Hoerl e Kennard (1970) introduz uma forma canônica ao modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ escrevendo $\mathbf{Y} = \mathbf{X}^*\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, em que $\mathbf{X}^* = \mathbf{XP}$ e $\boldsymbol{\alpha} = \mathbf{P}^t\boldsymbol{\beta}$ (SVD). Aqui, ϵ_i são independentes, possuem média 0 e variância constante.

As colunas de $\mathbf{X}^*$ são os $t = min(n, p)$ componentes de $\mathbf{X}$. Assim, o estimador de mínimos quadrados ordinários de $\boldsymbol{\alpha}$ é:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\Lambda}^{-1}\mathbf{X}^{*t}\mathbf{Y}, \tag{3.5}$$

que se relaciona com $\hat{\boldsymbol{\beta}}$ por: $\hat{\boldsymbol{\alpha}} = \mathbf{P}^t\hat{\boldsymbol{\beta}}$. Aqui, $\boldsymbol{\Lambda}$ e $\mathbf{P}$ são os autovalores e autovetores de $\mathbf{X}^t\mathbf{X}$, respectivamente, pela decomposição em valores singulares.

Em regressão por componentes principais, $\hat{\boldsymbol{\alpha}}$ são os coeficientes de regressão de um subconjunto dos componentes principais que formam as colunas de $\mathbf{X}^*$. Em regressão ridge, todos os coeficientes α_i são usados. Com o parâmetro k, estimativas dos coeficientes desta regressão ridge são dadas, na forma canônica, por:

$$\hat{\alpha}_{kj} = \frac{\lambda_j}{\lambda_j + k}\,\hat{\alpha}_j$$

Desta forma, Hoerl, Kennard e Baldwin (1975) propõe estimar k como:

$$k_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\alpha}}^t\hat{\boldsymbol{\alpha}}} = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^t\hat{\boldsymbol{\beta}}}$$

em que p é o número de covariáveis, $\hat{\boldsymbol{\alpha}}$ é calculado pela equação 3.5 e σ^2 é estimado por:

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p}.$$

Este estimador para k não é útil quando o número de preditores é maior do que o de observações, já que em tais situações os estimadores de mínimos quadrados ordinários de $\boldsymbol{\beta}$ e σ^2 não existem, resultando num k_{HKB} indefinido. Todavia, com o modelo em sua

forma canônica, k_{HKB} e σ^2 podem ser escritos como:

$$k_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\alpha}}^t\hat{\boldsymbol{\alpha}}} \quad \hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}^*\hat{\boldsymbol{\alpha}})^t(\mathbf{Y} - \mathbf{X}^*\hat{\boldsymbol{\alpha}})}{n-p}.$$

Desta forma, a proposta de Cule *et al.* (2012) é calcular k para uma quantidade pequena de componentes, que expliquem uma grande quantidade da variância de $\mathbf{X}$. Assim sendo, seu estimador é da forma:

$$k_a = \frac{a\,\hat{\sigma}_a^2}{\sum_{j=1}^{a}\hat{\alpha}_j^2}$$

em que a é o número de componentes selecionados ($a < p$) e $\hat{\sigma}_a^2$ é dado por:

$$\hat{\sigma}_a^2 = \frac{(\mathbf{y} - \mathbf{X}_a^*\hat{\boldsymbol{\alpha}}_a)^t(\mathbf{y} - \mathbf{X}_a^*\hat{\boldsymbol{\alpha}}_a)}{n-a}$$

em que $\mathbf{X}_a^*$ e $\hat{\boldsymbol{\alpha}}_r$ contém as primeiras a colunas e os primeiros a elementos de $\mathbf{X}^*$ e $\hat{\boldsymbol{\alpha}}$, respectivamente. Este estimador, com base nas simulações de Cule *et al.* (2012), apresentou menores erros quadráticos médios do que o estimador de Hoerl, Kennard e Baldwin (1975) para casos em que ($p < n$) e os erros de regressão não são tão pequenos.

Para a escolha do número de componentes a ser utilizado, pode ser utilizada uma validação cruzada com base na estatística PRESS (De Iorio, Ebbles e Stephens, 2007), que é soma dos quadrados dos resíduos preditivos, da mesma forma que é feito em mínimos quadrados parciais (PLS).

3.4 LASSO

Modelos de regressão LASSO (Tibshirani, 1996) são alternativas de modelagem para casos em que há muitas covariáveis, e em que a relação destas com a variável resposta não é tão clara. Basicamente, o LASSO minimiza a soma dos quadrados dos resíduos sujeito a restrição de que a soma dos valores absolutos dos coeficientes é menor do que uma constante. Por causa da natureza desta restrição, a técnica tende a produzir algumas estimativas de coeficientes que são exatamente iguais a 0, o que não ocorre na regressão ridge, fazendo com que o modelo selecionado seja mais interpretável, com menor número de covariáveis.

De uma forma geral, considere a situação usual em que temos um conjunto de dados $(\mathbf{x}^i,\ y_i)$ para $i = 1,\ 2,\ ...,\ \text{n}$, em que $\mathbf{x}^i = (x_{i1}, x_{i2}, ..., x_{ip})^t$ e y_i são as covariáveis e a variável resposta para a i-ésima observação, respectivamente. As estimativas para os coeficientes de regressão via mínimos quadrados ordinários são obtidas minimizando a soma dos quadrados dos erros. Entretanto, há duas razões principais que fazem com que este tipo de estimador não seja satisfatório para a análise desses dados. A primeira razão é a questão da acurácia das predições: os estimadores de mínimos quadrados podem ter, em alguns casos, vício pequeno, mas alta variância. Assim sendo, aumentando um pouco o vício para reduzir a variância dos valores preditos, podemos melhorar a previsão geral, ou seja, a acurácia, que é o que o LASSO faz. A segunda razão é a questão da interpretação: com um grande número de preditores, torna-se difícil interpretar todos os coeficientes estimados. Desta forma, o LASSO exibiria apenas os efeitos mais fortes e importantes, estimando os demais como iguais a zero.

As estimativas do LASSO são obtidas minimizando a seguinte expressão:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad sujeito\ a \quad \sum_{j=1}^{p}|\beta_j| \leq t, \tag{3.6}$$

em que $t \geq 0$ é um parâmetro de ajuste na qual para todo valor que t assume, a solução para β_0 é $\hat{\beta}_0 = \overline{y}$. Além disso, na expressão 3.6 é assumido que as covariáveis são centradas e escalonadas.

O parâmetro $t \geq 0$ controla o "valor de encolhimento" que é aplicado para as estimativas. Seja $\hat{\beta}_j^o$ as estimativas de mínimos quadrados ordinários e seja $t_0 = \sum_{j=1}^{p}\left|\hat{\beta}_j^o\right|$. Valores de $t < t_0$ irão causar o "encolhimento" das soluções para 0, com alguns coeficientes estimados podendo ser exatamente iguais a 0. Por exemplo, se $t = t_0/2$, o efeito será mais ou menos semelhante a encontrar o melhor subconjunto de covariáveis de tamanho $p/2$.

3.4.1 Erro de predição e escolha de t

Nesta seção é descrito um método para a escolha da constante t, que é detalhado por Hastie *et al.* (2009).

Suponha que temos o modelo $\text{Y} = \eta(\mathbf{X}) + \epsilon$, em que $\text{E}(\epsilon) = 0$ e a $\text{var}(\epsilon) = \sigma^2$. O erro quadrático médio de um estimador $\hat{\eta}(\mathbf{X})$ é definido por $ME = E\left[\hat{\eta}(\mathbf{X}) - \eta(\mathbf{X})\right]^2$. Assim,

podemos calcular o erro de predição de $\hat{\eta}(\mathbf{X})$ por:

$$PE = E\left[Y - \hat{\eta}(\mathbf{X})\right]^2 = ME + \sigma^2.$$

Desta forma, a estimação do erro de predição pode ser utilizada como um critério de seleção dentro de uma validação cruzada, como utilizada para as técnicas anteriores. Nesta abordagem, o LASSO é indexado em termos do parâmetro $s = t/\sum_{j=1}^{p}\left|\hat{\beta}_j^o\right|$, e o erro de predição é estimado sobre uma grade de valores de s dentro do intervalo [0,1]. O valor de $\hat{s}$ que obter o menor PE é selecionado, encontrando, por consequência, o corresponde valor de t.

A figura 3.2 representa a variação dos valores dos coeficientes estimados via LASSO para cada valor de s, em um exemplo de Hastie *et al.* (2009). O valor com menor PE, encontrado no trabalho do autor e calculado por meio de uma validação cruzada, foi de 0.36.

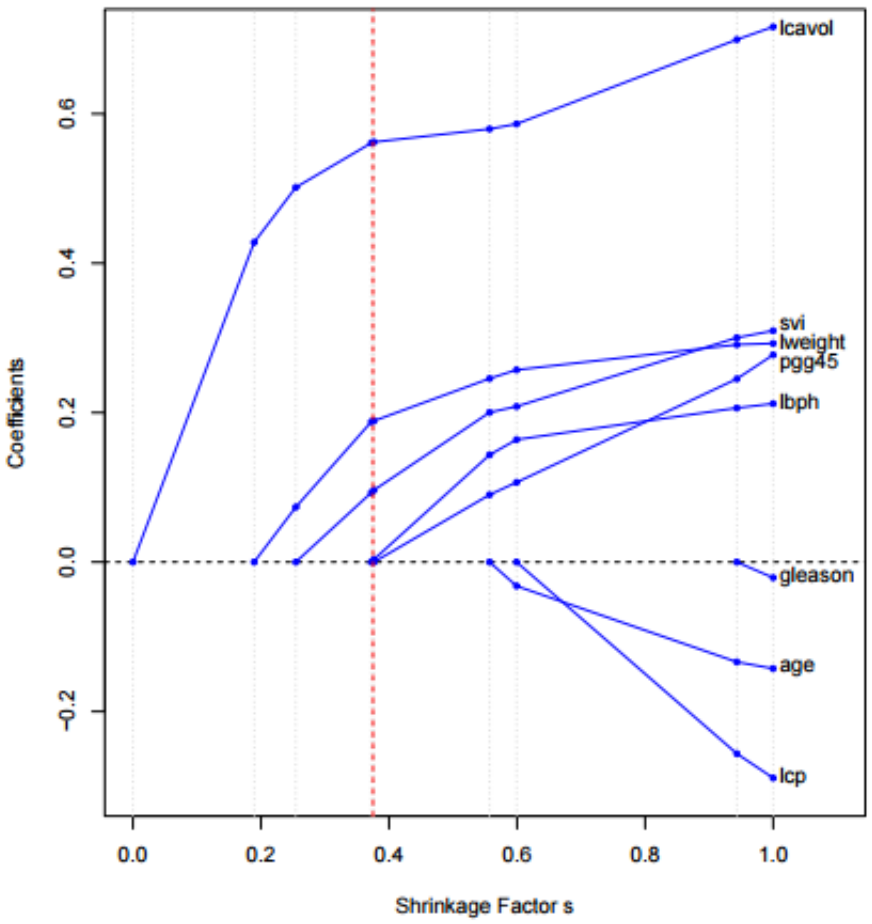

Figura 3.2: Validação cruzada com base no erro de predição (PE) para cada valor de s (Hastie, 2009)

No próximo capítulo serão apresentadas as simulações realizadas com o intuito de comparar as quatros técnicas discutidas até aqui com relação ao poder de predição delas.

Capítulo 4

Estudo de Simulação

Neste capítulo apresentamos o estudo comparativo do poder de predição das quatro técnicas de regressão estudadas: componentes principais (PCR), mínimos quadrados parciais (PLS), regressão ridge (RR) e LASSO para diferentes cenários de tamanho de amostra, número de covariáveis e quantidade e intensidade de coeficientes (efeitos) significativos, abordando os problemas de multicolinearidade e de alta dimensionalidade, em que $p > n$, que podem causar resultados instáveis nos métodos de regressão por mínimos quadrados ordinários, além do aumento da variância dos coeficientes estimados.

As simulações são divididas em duas partes. Na primeira, a intensidade e a quantidade de efeitos significativos é levada em consideração, em um caso em que os dados contêm sérios problemas de multicolinearidade, mas o número de covariáveis é menor do que o número de observações. Neste contexto são estudados os seguintes cenários, divididos em exemplos:

- **Grande número de pequenos efeitos.** Situação na qual há vários coeficientes significativos, mas de intensidade baixa;

- **Grande número de grandes efeitos.** Situação na qual há vários coeficientes significativos e com intensidade relativamente alta;

- **Pequeno número de grandes efeitos.** Situação na qual há poucos coeficientes significativos, mas com intensidade relativamente alta;

- **Moderado número de moderados efeitos.** Situação na qual boa parte dos coeficientes são significativos, e a intensidade dos mesmos é moderada;

- **Variação na intensidade dos efeitos, com porcentagem de efeitos não-nulos.** Situação dividida em três cenários, na qual diferentes porcentagens de efeitos não-nulos são levadas em consideração em casos em que há variação na intensidade desses efeitos, ou seja, há pequenos, moderados e grandes efeitos ao mesmo tempo.

Na segunda parte os problemas de multicolinearidade e de alta dimensionalidade, para $p > n$, são abordados. Neste contexto, dois diferentes valores de tamanho de amostra ($n =$ 30, 50) são fixados e avaliados para oito diferentes quantidades de número de covariáveis. Para o caso em que $n = 30$, é considerado que $p = 30$, 60, 90 e 120. Já para o caso em que $n = 50$, é considerado que $p = 50$, 100, 150 e 200.

4.1 Métodos e suporte computacional

Antes de apresentarmos os resultados das simulações, destacaremos os métodos desenvolvidos e os procedimentos realizados.

Para a escolha do número de componentes a ser utilizado nas regressões por componentes principais (PCR) e nas regressões por mínimos quadrados parciais (PLS) foram usadas validações cruzadas com base na proposta de Mertens *et al.* (1995). Assim sendo, foi utilizada a biblioteca *pls* do *software* R, criando esses modelos pelos comandos *pcr()* e *plsr()*, respectivamente.

Na estimação da constante da regressão ridge (RR) foi usada a metodologia de Cule *et al.* (2012), destacada anteriormente. Dessa forma, foi utilizada a biblioteca *ridge* do *software* R, criando os modelos ridge pelo comando *linearRidge()*.

Já para a estimação do parâmetro s, que indexa o LASSO, foi usada a abordagem apresentada no capítulo anterior. Computacionalmente, os modelos LASSO foram criados utilizando o comando *lars()* da biblioteca *lars*, que também é encontrada no *software* R.

4.2 Parte I

Nesta seção são apresentados os resultados obtidos para diferentes cenários de quantidade e intensidade de efeitos significativos presentes na regressão.

Para os cenários que serão apresentados na Parte I, 25 covariáveis foram geradas de uma distribuição normal multivariada com vetor de médias igual a $\mathbf{0}$ e matriz de variâncias-covariâncias dada por $\mathbf{\Sigma}_{i,j} = 10 \times 0.9^{(|i-j|)}$. Estas covariáveis são consideradas

fixas dentro de cada uma das B = 500 replicações das simulações que serão desenvolvidas. Os erros, para cada uma das $n = 100$ observações criadas, são gerados de uma distribuição normal com média 0 e variância 4, e se alteram para cada uma das replicações. Já as respostas são calculadas com base na equação: $y_i = \beta_0 + \beta_1 x_1 + ... + \beta_{25} x_{25} + \epsilon_i$, na qual os valores dos coeficientes; β_0, β_1, ..., β_{25}; são fixados.

A tabela a seguir apresenta os valores aproximados de VIF (Variance Inflation Factor) das covariáveis com base em um possível ajuste de um modelo de regressão por mínimos quadrados ordinários. O VIF_k (VIF relacionado a k-ésima covariável) mede o quanto a variância do coeficiente da regressão padronizada é inflacionada por sua colinearidade e é um dos principais indicadores de problemas de colinearidade usados na literatura. De uma forma geral, VIFs maiores do que 10 indicam sérios problemas de colinearidade.

Tabela 4.1: VIF_k das covariáveis

X1	X2	X3	X4	X5
10.337	13.667	10.598	14.044	9.988
X6	X7	X8	X9	X10
14.163	15.303	11.098	15.256	20.248
X11	X12	X13	X14	X15
27.678	19.876	26.414	24.072	23.272
X16	X17	X18	X19	X20
19.327	10.478	16.131	17.508	19.048
X21	X22	X23	X24	X25
16.887	21.644	22.787	13.835	13.042

Analisando a Tabela 4.1 observamos que há sérios problemas de multicolinearidade, com quase todos os VIFs maiores do que 10. A média deles é de, aproximadamente, 17.

4.2.1 Cenário 1

No primeiro cenário é considerado o caso em que há grande número de pequenos efeitos. Assim sendo, é considerado que $\beta_0 = 1$, $\beta_1 = 0.85$, $\beta_2 = 0.85$, ..., $\beta_{25} = 0.85$. Para a obtenção dos resultados, a amostra de tamanho $n = 100$ é dividida em amostra treinamento, contendo 70% dos dados (70 observações), e em amostra teste, contendo os

30% restantes (30 observações). Os quatro modelos são ajustados com base na amostra treinamento e comparados, segundo os seus poderes de predição, com base na amostra teste.

A tabela 4.2 apresenta a mediana das B = 500 médias dos resíduos ao quadrado obtidas para cada modelo na predição das observações da amostra teste e seus respectivos desvios padrões, em parênteses. Além disso, a tabela demonstra a proporção de vezes dentre as B = 500 replicações que cada modelo obteve a menor Distância Relativa (DR), que é dada por:

$$DR = \frac{1}{N_{teste}} \sum_{i=1}^{N_{teste}} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100, \quad Y_i \neq 0,$$

em que $N_{teste} = 30$ é o tamanho da amostra teste, Y_i é a i-ésima resposta da amostra teste e $\hat{Y}_i$ é o valor predito para Y_i segundo o modelo estimado. Quanto menor o valor obtido, melhor é a predição do método.

Tabela 4.2: Resultados do cenário 1 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	5.996 (2.346)	0.162
PLS	5.907 (2.316)	0.158
RR	5.436 (2.095)	0.592
LASSO	6.051 (2.356)	0.088

Analisando a Tabela 4.2 observamos que a regressão ridge apresentou os melhores resultados, com menor mediana e desvio padrão (variabilidade) dentre as quatro técnicas. Com relação à DR, a regressão ridge também foi a que mais se destacou, obtendo melhores predições em 59.2% das vezes.

Uma outra informação relevante em estudos comparativos de predição é o tempo computacional gasto para que o *software* utilizado possa criar o modelo e realizar as predições da(s) variável(is) resposta das novas observações. Assim sendo, a tabela a seguir apresenta a média do tempo computacional gasto para cada umas das B = 500 replicações, avaliada em segundos, com base no cenário 1.

Tabela 4.3: Resultados do cenário 1 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.032
PLS	0.045
RR	0.197
LASSO	0.102

Analisando a Tabela 4.3 observamos que, apesar da medida calculada depender do computador utilizado, a regressão ridge foi a que obteve o pior resultado, gastando um tempo computacional médio maior do que os demais métodos. Os procedimentos mais rápidos foram os desenvolvidos para as regressões por componentes principais.

Já a Tabela 4.4, logo a seguir, apresenta a proporção de vezes que o coeficiente associado à cada covariável foi estimado como diferente de 0 dentro das B = 500 replicações realizadas, com base no LASSO. Dessa forma, observamos que os resultados foram satisfatórios, pois todos os coeficientes foram fixados como diferentes de 0 e o LASSO conseguiu captar essa característica com grande eficiência.

Tabela 4.4: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	1	1	0.996	0.996
X6	X7	X8	X9	X10
1	0.998	1	1	0.998
X11	X12	X13	X14	X15
0.996	0.998	0.994	0.998	1
X16	X17	X18	X19	X20
0.998	1	0.994	0.998	0.998
X21	X22	X23	X24	X25
1	1	0.988	1	1

4.2.2 Cenário 2

No segundo cenário é considerado o caso em que há grande número de grandes efeitos. Assim sendo, é considerado que $\beta_0 = 1$, $\beta_1 = 5$, $\beta_2 = 5$, ..., $\beta_{25} = 5$. Os demais procedimentos, citados no cenário anterior, foram mantidos.

Analisando a Tabela 4.5 com base na mediana das médias dos resíduos ao quadrado, observamos que nenhuma das técnicas se sobressaiu perante as demais, com resultados bem parecidos. Entretanto, quando analisamos a DR, concluímos que a regressão ridge ainda obteve o maior poder de predição.

Tabela 4.5: Resultados do cenário 2 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	6.021 (2.341)	0.208
PLS	5.910 (2.307)	0.318
RR	5.964 (2.318)	0.416
LASSO	6.035 (2.350)	0.058

Já quando analisamos a Tabela 4.6 observamos que a regressão ridge foi a que obteve o pior resultado, gastando um tempo computacional médio maior do que os demais métodos. Além disso, assim como no cenário 1, os procedimentos mais rápidos foram os desenvolvidos para as regressões por componentes principais.

Tabela 4.6: Resultados do cenário 2 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.032
PLS	0.047
RR	0.204
LASSO	0.107

A tabela 4.7, associada aos coeficientes estimados utilizando o LASSO, demonstra que essa metodologia apresentou excelentes resultados, à medida que todos os coeficientes que

foram fixados como iguais a 5 foram incluídos nos modelos em todas as replicações do algoritmo.

Tabela 4.7: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	1	1	1	1
X6	X7	X8	X9	X10
1	1	1	1	1
X11	X12	X13	X14	X15
1	1	1	1	1
X16	X17	X18	X19	X20
1	1	1	1	1
X21	X22	X23	X24	X25
1	1	1	1	1

4.2.3 Cenário 3

No terceiro cenário é considerado o caso em que há pequeno número de grandes efeitos. Assim sendo, é considerado que $\beta_0 = 5$, $\beta_1 = 10$, $\beta_2 = 10$, $\beta_3 = 10$, $\beta_4 = 0$, $\beta_5 = 0$,..., $\beta_{25} = 0$. Os demais procedimentos, citados no primeiro cenário, foram mantidos.

Analisando a Tabela 4.8 observamos que o LASSO apresentou os melhores resultados por uma boa margem, com menor mediana e desvio padrão dentre as quatro técnicas. Isto fica bem claro, também, com os resultados encontrados para a DR, em que o LASSO obteve melhores predições em 76% das replicações.

Tabela 4.8: Resultados do cenário 3 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	6.028 (2.345)	0.052
PLS	5.911 (2.315)	0.140
RR	6.022 (2.339)	0.048
LASSO	4.294 (1.640)	0.760

Com relação ao tempo computacional médio gasto por cada método e apresentado pela Tabela 4.9, é possível destacar que os mesmos resultados encontrados para os dois primeiros cenários foram mantidos. A regressão ridge foi a que obteve o pior resultado, enquanto que a regressão por componentes principais foi a que despendeu menos tempo.

Tabela 4.9: Resultados do cenário 3 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.036
PLS	0.047
RR	0.213
LASSO	0.115

Já a tabela associada aos coeficientes estimados utilizando o LASSO, apresentada logo a seguir, destaca como a metodologia conseguiu captar os efeitos diferentes de 0. Apenas os três primeiros coeficientes foram incluídos em todas as replicações do algoritmo, e justamente tais coeficientes eram diferentes de 0. Os demais, que eram iguais a 0, tiveram proporções abaixo de 0.420.

Tabela 4.10: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	1	1	0.414	0.372
X6	X7	X8	X9	X10
0.342	0.314	0.382	0.322	0.290
X11	X12	X13	X14	X15
0.288	0.298	0.272	0.246	0.246
X16	X17	X18	X19	X20
0.310	0.382	0.308	0.308	0.338
X21	X22	X23	X24	X25
0.310	0.278	0.290	0.336	0.352

4.2.4 Cenário 4

No quarto cenário é considerado o caso em que há moderado número de moderados efeitos. Assim sendo, é considerado que $\beta_0 = 1$, $\beta_1 = 2$, ..., $\beta_5 = 2$, $\beta_6 = 0$, ..., $\beta_{10} = 0$, $\beta_{11} = 2$, ..., $\beta_{15} = 2$, $\beta_{16} = 0$, ..., $\beta_{20} = 0$, $\beta_{21} = 2$, ..., $\beta_{25} = 2$. Os demais procedimentos, citados no primeiro cenário, foram mantidos.

Analisando a Tabela 4.11 observamos que o LASSO apresentou os melhores resultados, com menor mediana, menor desvio padrão e maior proporção de vezes em que a técnica apresentou menor DR (51.4%). A regressão ridge obteve os segundos melhores valores.

Tabela 4.11: Resultados do cenário 4 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	6.046 (2.347)	0.164
PLS	6.046 (2.367)	0.112
RR	5.885 (2.236)	0.210
LASSO	5.552 (2.164)	0.514

Analisando a Tabela 4.12 observamos que as mesmas características encontradas nos cenários anteriores foram mantidas. A regressão por componentes principais continuou sendo o método mais rápido.

Tabela 4.12: Resultados do cenário 4 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.032
PLS	0.046
RR	0.199
LASSO	0.114

Já com relação aos coeficientes estimados utilizando o LASSO, observamos que todos os coeficientes fixados como iguais a 2 foram incluídos nos modelos em todas as 500 replicações realizadas. Já os coeficientes que eram iguais a 0 obtiveram proporções de inclusão menores, abaixo de 0.680.

Tabela 4.13: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	1	1	1	1
X6	X7	X8	X9	X10
0.670	0.534	0.618	0.566	0.590
X11	X12	X13	X14	X15
1	1	1	1	1
X16	X17	X18	X19	X20
0.650	0.574	0.534	0.512	0.578
X21	X22	X23	X24	X25
1	1	1	1	1

4.2.5 Cenário 5

Neste e nos próximos dois últimos cenários são considerados os casos em que há variação na intensidade dos efeitos, ou seja, há pequenos, moderados e grandes efeitos ao mesmo tempo, em um único modelo. Esta situação é a mais comum na prática, à medida que dificilmente se encontrará um modelo em que os coeficientes possuem valores iguais ou até mesmo próximos entre si. Com relação à quantidade, este cenário apresentará 20% de efeitos não nulos. Já os cenários 6 e 7, que virão logo em seguida, apresentarão 40 e 60%, respectivamente.

Assim sendo, para este cenário, é considerado que o vetor $\boldsymbol{\beta}$, incluindo o intercepto, é da seguinte forma: (1, 3, 0, -4, 0, 0, 0, 0, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0, -9, 0, 0, 0, 0, 6).

Analisando a Tabela 4.14 observamos que o LASSO obteve os melhores resultados com base em todos os aspectos analisados, possuindo a menor DR em 68.8% das replicações.

Tabela 4.14: Resultados do cenário 5 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	6.035 (2.350)	0.072
PLS	6.033 (2.353)	0.108
RR	6.072 (2.317)	0.132
LASSO	4.493 (1.700)	0.688

Com relação ao tempo computacional médio gasto por cada técnica, observamos que as mesmas características encontradas nos cenários anteriores foram mantidas. A regressão por componentes principais continuou sendo o método mais rápido, gastando cerca de 0.033 segundos para modelar e predizer em cada uma das 500 replicações.

Tabela 4.15: Resultados do cenário 5 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.033
PLS	0.045
RR	0.197
LASSO	0.105

Já com base nos coeficientes estimados utilizando o LASSO, observamos que todos os coeficientes não-nulos obtiveram proporções iguais a 1, ou seja, foram incluídos nos modelos criados em todas as 500 replicações. Enquanto isso, os efeitos nulos obtiveram proporções de inclusão abaixo de 0.525.

Tabela 4.16: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	0.346	1	0.348	0.408
X6	X7	X8	X9	X10
0.384	0.330	0.524	0.326	1
X11	X12	X13	X14	X15
0.380	0.366	0.320	0.268	0.302
X16	X17	X18	X19	X20
0.354	0.398	0.352	0.380	1
X21	X22	X23	X24	X25
0.394	0.294	0.328	0.346	1

4.2.6 Cenário 6

Como já comentado, este cenário considera o caso em que 40% dos coeficientes são não nulos. Desta forma, é considerado que o vetor $\boldsymbol{\beta}$, incluindo o intercepto, é da seguinte forma: (1, 3, 5, -4, 0, 0, 2, 0, 0, 0, 8, 0, 0, -5, 0, 0, 1, 0, 0, 3, 9, 0, 0, 0, 0, 1).

Analisando a Tabela 4.17 observamos que o LASSO ainda obteve os melhores resul-

tados com base em todos os aspectos analisados. Entretanto, sua eficiência foi menor do que para o caso anterior.

Tabela 4.17: Resultados do cenário 6 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	6.035 (2.350)	0.294
PLS	6.046 (2.353)	0.222
RR	5.975 (2.399)	0.108
LASSO	5.301 (2.053)	0.376

Com relação ao tempo computacional médio gasto por cada técnica, observamos que as mesmas características encontradas nos cenários anteriores foram mantidas. A regressão por componentes principais continuou sendo o método mais rápido, gastando cerca de 0.032 segundos para modelar e predizer em cada uma das 500 replicações.

Tabela 4.18: Resultados do cenário 6 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.032
PLS	0.045
RR	0.199
LASSO	0.119

Já com base nos coeficientes estimados utilizando o LASSO, observamos que quase todos os coeficientes não-nulos obtiveram proporções iguais a 1. O único caso que isto não ocorreu foi para a covariável 16, que foi fixada como igual a 1 e obteve proporção de 0.998. Enquanto isso, os efeitos nulos obtiveram proporções de inclusão abaixo de 0.675.

Tabela 4.19: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	1	1	0.476	0.622
X6	X7	X8	X9	X10
1	0.668	0.570	0.534	1
X11	X12	X13	X14	X15
0.386	0.556	1	0.426	0.590
X16	X17	X18	X19	X20
0.998	0.586	0.670	1	1
X21	X22	X23	X24	X25
0.626	0.488	0.586	0.488	1

4.2.7 Cenário 7

No último cenário é considerado o caso em que 60% dos coeficientes são não nulos. Desta forma, é considerado que o vetor $\boldsymbol{\beta}$, incluindo o intercepto, é da seguinte forma: (1, 3, 5, -4, 0, 0, 2, 0, -7, 0, 8, 0, 0, -5, 4, 0, 1, 6, 0, 3, 9, 0, 2, -8, 0, 1).

Analisando a Tabela 4.20 observamos que a tendência percebida no cenário anterior se manteve, ou seja, apesar do LASSO ainda obter os melhores resultados, sua eficiência continuou a diminuir.

Tabela 4.20: Resultados do cenário 7 - Medidas preditivas

Método	Mediana das médias dos resíduos ao quadrado	Proporção de vezes em que o método apresentou menor DR
PCR	6.035 (2.350)	0.202
PLS	6.031 (2.351)	0.356
RR	5.967 (2.400)	0.082
LASSO	5.861 (2.298)	0.360

Com relação ao tempo computacional médio gasto por cada técnica, observamos que as mesmas características encontradas nos cenários anteriores foram mantidas. A re-

gressão ridge baseada na proposta de Cule *et al.* (2012) continuou sendo o método mais lento, gastando cerca de 0.197 segundos para modelar e predizer em cada uma das 500 replicações.

Tabela 4.21: Resultados do cenário 7 - Tempo computacional

Método	Tempo computacional médio (em segundos)
PCR	0.032
PLS	0.046
RR	0.197
LASSO	0.115

Já com base nos coeficientes estimados utilizando o LASSO, observamos que quase todos os coeficientes não-nulos obtiveram proporções iguais a 1. O único caso que isto não ocorreu foi para a covariável 16, que foi fixada como igual a 1 e obteve proporção de 0.998. Enquanto isso, os efeitos nulos obtiveram proporções de inclusão abaixo de 0.810.

Uma característica que foi possível destacar, analisando apenas os três últimos cenários, foi que com o aumento da proporção de efeitos não-nulos considerados (20, 40 e 60%), as proporções de vezes que o LASSO considerou cada coeficiente nulo como diferente de zero também aumentaram.

Tabela 4.22: Proporção de vezes que o modelo LASSO considerou cada coeficiente como diferente de 0 dentro das B = 500 replicações das simulações

X1	X2	X3	X4	X5
1	1	1	0.722	0.808
X6	X7	X8	X9	X10
1	0.764	1	0.728	1
X11	X12	X13	X14	X15
0.632	0.724	1	1	0.756
X16	X17	X18	X19	X20
0.998	1	0.778	1	1
X21	X22	X23	X24	X25
0.788	1	1	0.756	1

4.2.8 Comentários gerais - Parte I

Analisando os resultados obtidos com base apenas nas questões preditivas, foi possível destacar os seguintes padrões:

- Quando houve um grande número de pequenos efeitos, a regressão ridge foi a melhor opção;

- Quando houve um grande número de grandes efeitos, a regressão ridge também foi a melhor opção, mas tanto o PLS quanto o PCR não obtiveram resultados tão ruins;

- Quando houve um pequeno número de grandes efeitos, o LASSO foi a melhor opção;

- Quando houve um moderado número de moderados efeitos, o LASSO também foi a melhor opção;

- Quando houve variação na intensidade dos efeitos, ou seja, houve pequenos, moderados e grandes efeitos ao mesmo tempo, o LASSO foi a melhor opção dependendo da quantidade de efeitos não nulos que teve. Quanto menor foi essa quantidade, melhores foram os resultados obtidos pelo LASSO se comparado as outras três técnicas.

Com relação ao tempo computacional médio, gasto por cada técnica para modelar e estimar novas observações, a regressão por componentes principais obteve os melhores resultados, sendo a técnica mais rápida em todos os cenários. Enquanto isso, a regressão ridge foi a que apresentou os piores resultados, sendo o método mais lento dentre os quatro.

Já com relação aos coeficientes estimados pelo LASSO, observamos que a técnica apresentou excelente qualidade para identificar os coeficientes significativos, considerando os mesmos em quase 100% dos casos. Entretanto, quando a proporção de efeitos não-nulos foi aumentando, a técnica apresentou uma maior tendência a incluir em seus modelos covariáveis nulas, que não seriam relevantes teoricamente. Isto foi possível observar visualizando as proporções de inclusão associadas aos coeficientes que foram fixados como iguais a 0 e passaram a ser inclusos nos modelos com maiores frequências, à medida que a quantidade de efeitos significativos cresceu.

4.3 Parte II

Nesta seção são apresentados os resultados obtidos para os casos de alta dimensionalidade em que o número de covariáveis é maior do que o de observações e há problemas de multicolinearidade.

Para os três cenários que serão apresentados, são considerados os casos dos cenários 5, 6 e 7 da Parte I, em que há variação na intensidade dos efeitos em um único modelo, e em três diferentes situações: com 20, 40 e 60% de efeitos não nulos. O número de covariáveis considerado foi fixado em quatro diferentes valores para cada tamanho de amostra pré-determinado. Assim sendo, para $n = 30$ foi estabelecido $p = 30$, 60, 90 e 120. Já para $n = 50$ foi estabelecido $p = 50$, 100, 150 e 200. Os demais procedimentos, citados anteriormente, foram mantidos, ou seja, as covariáveis e os erros foram gerados da mesma forma, e as respostas foram calculadas da mesma maneira. A divisão em 70 e 30%, para a criação das amostras treinamento e teste, respectivamente, também é utilizada.

4.3.1 Cenário 1

A Tabela 4.23 apresenta os resultados obtidos para o caso em que há 20% de efeitos não nulos. Analisando-a, observamos que não houve nenhum padrão nos resultados. Para cada situação, um determinado modelo foi melhor do que o outro.

Quando o n foi igual a 30 e o p foi igual a 60, por exemplo, o LASSO obteve os melhores resultados em 94.2% das replicações. Já quando o n foi igual a 50 e o p foi igual a 100, ou seja, no caso em que a proporção entre o número de covariáveis e o número de observações, com base no cenário, foi mantida, a regressão ridge se destacou mais, seguida da regressão por componentes principais.

Tabela 4.23: Proporção da DR (20% de efeitos não nulos)

$n = 30$	PCR	PLS	RR	LASSO	$n = 50$	PCR	PLS	RR	LASSO
$p = 30$	0.028	0.692	0.054	0.226	$p = 50$	0.002	0	0	0.998
$p = 60$	0.024	0.030	0.004	0.942	$p = 100$	0.404	0	0.596	0
$p = 90$	0	0	1	0	$p = 150$	0	0	0.592	0.408
$p = 120$	0	0	0.964	0.036	$p = 200$	0.814	0	0.004	0.182

4.3.2 Cenário 2

A Tabela 4.24 apresenta os resultados obtidos para o caso em que há 40% de efeitos não nulos. Analisando-a, observamos que não houve nenhum padrão nos resultados. Para cada situação, um determinado modelo foi melhor do que o outro.

Por exemplo, quando o n foi igual a 30 e o p foi igual a 60, o LASSO obteve os melhores resultados em 62.8% das replicações. Já quando o n foi igual a 50 e o p foi igual a 100, o PLS foi praticamente soberano, obtendo os melhores resultados em 99.6% das vezes.

Além disso, houve quatro casos em que um dos modelos foi considerado o melhor em 100% das replicações. Quando o n foi igual a 30 e o p foi igual a 90, e quando o n foi igual a 50 e o p foi igual a 200, o LASSO foi a melhor opção. Já quando o n foi igual a 30 e o p foi igual a 120, e quando o n foi igual a 50 e o p foi igual a 150, o PLS foi a melhor escolha.

Tabela 4.24: Proporção da DR (40% de efeitos não nulos)

$n = 30$	PCR	PLS	RR	LASSO	$n = 50$	PCR	PLS	RR	LASSO
$p = 30$	0.104	0.070	0.004	0.822	$p = 50$	0.014	0.020	0.006	0.960
$p = 60$	0.222	0	0.150	0.628	$p = 100$	0	0.996	0	0.004
$p = 90$	0	0	0	1	$p = 150$	0	1	0	0
$p = 120$	0	1	0	0	$p = 200$	0	0	0	1

4.3.3 Cenário 3

A Tabela 4.25 apresenta os resultados obtidos para o caso em que há 60% de efeitos não nulos. Analisando-a, observamos que não houve nenhum padrão nos resultados. Para cada situação, um determinado modelo foi melhor do que o outro.

Por exemplo, quando o n foi igual a 30 e o p foi igual a 60, o LASSO obteve os melhores resultados em 70.4% das replicações. Já quando o n foi igual a 50 e o p foi igual a 100, o PCR foi soberano, obtendo os melhores resultados em todas as replicações.

Tabela 4.25: Proporção da DR (60% de efeitos não nulos)

$n = 30$	PCR	PLS	RR	LASSO	$n = 50$	PCR	PLS	RR	LASSO
$p = 30$	0.132	0.462	0.108	0.298	$p = 50$	0.054	0.010	0.936	0
$p = 60$	0	0	0.296	0.704	$p = 100$	1	0	0	0
$p = 90$	0.226	0	0.596	0.178	$p = 150$	0	0.998	0	0.002
$p = 120$	0	1	0	0	$p = 200$	0	0	0	1

4.3.4 Comentários gerais - Parte II

Analisando os resultados obtidos para a segunda parte, em que o número de covariáveis se torna maior do que o número de observações e os problemas de multicolinearidade foram mantidos, pudemos destacar que não houve nenhum padrão. Para cada situação, um determinado modelo se destacou mais.

Assim sendo, o próximo capítulo apresentará uma aplicação associada a exemplos com essas características, ou seja, em que há alta dimensionalidade, com maior número de covariáveis do que de observações, e em que há sérios problemas de colinearidade.

Capítulo 5

Aplicação em Dados Reais

Para ilustrar os casos de alta dimensionalidade, em que o número de covariáveis é maior do que o número de observações, as quatro técnicas estudadas são comparadas em um conjunto de dados reais (não simulados).

O conjunto de dados em questão é aquele que foi discutido no primeiro capítulo, encontrado na biblioteca *pls* do *software* R e sendo carregado utilizando o comando *data(gasoline)*. Este banco de dados apresenta 60 observações e fornece o índice de octano de amostras de gasolina junto com outras 401 variáveis que medem o índice de espectros NIR (espectroscopia de infravermelho próximo) dessas amostras. O interesse é predizer o índice de octano (variável resposta) pelos índices de espectros NIR dessas 401 medidas de refletância.

Para a obtenção dos resultados que serão apresentados logo a seguir, a amostra de tamanho $n = 60$ foi dividida em amostra treinamento, contendo 70% dos dados (42 observações), e em amostra teste, contendo os 30% restantes (18 observações). Os quatro modelos são ajustados com base na amostra treinamento e comparados, segundo os seus poderes de predição, com base na amostra teste.

5.1 Resultados

Inicialmente apresentamos a Tabela 5.1, que destaca a média dos resíduos ao quadrado obtida para cada modelo na predição das observações da amostra teste e os respectivos desvios padrões, em parênteses, desses resíduos. Além disso, a tabela demonstra a distância relativa obtida para cada modelo.

Tabela 5.1: Resultados da aplicação - Medidas preditivas

Método	Média dos resíduos ao quadrado	Distância Relativa (DR)
PCR	0.048 (0.038)	0.223
PLS	0.057 (0.047)	0.241
RR	0.048 (0.046)	0.215
LASSO	0.068 (0.068)	0.257

Analisando os resultados obtidos observamos que a regressão por componentes principais e a regressão ridge, baseada na proposta de Cule *et al.* (2012), foram as técnicas que obtiveram os melhores resultados com relação à média dos resíduos ao quadrado. Entretanto, quando olhamos o desvio padrão, a regressão por componentes principais foi o método que obteve o menor valor dentre as quatro opções. Quanto as distâncias relativas, a regressão ridge foi a melhor abordagem, com um valor de 0.215.

A figura a seguir apresenta os boxplots dos resíduos ao quadrado. Analisando-a podemos observar que, apesar do desvio padrão da regressão por componentes principais ter sido menor, os resultados da regressão ridge possuíram menor variabilidade, descontado um *outlier* que se tornou presente.

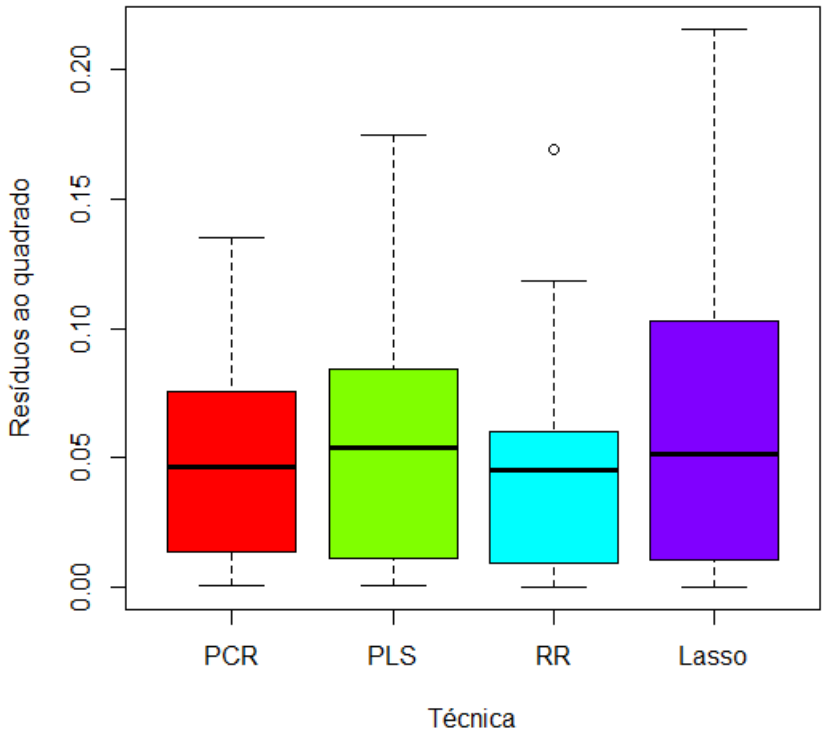

Figura 5.1: Boxplots dos resíduos ao quadrado para as quatro técnicas abordadas na aplicação

Com base em todas as medidas analisadas, o LASSO, que selecionou 17 covariáveis, foi a pior metodologia. A média dos resíduos ao quadrado, associada à esta metodologia, foi de 0.068, assim como seu desvio padrão. Já a distância relativa foi de 0.257, maior valor dentre as quatro técnicas abordadas.

A tabela a seguir apresenta o tempo computacional gasto (medido em segundos) para que o *software* utilizado conseguisse criar o modelo e realizar as estimações da variável resposta das novas observações.

Tabela 5.2: Resultados da aplicação - Tempo computacional

Método	Tempo computacional (em segundos)
PCR	0.406
PLS	0.531
RR	0.719
LASSO	2.219

Analisando a Tabela 5.2 observamos que, apesar da medida calculada depender do computador utilizado, o LASSO foi a técnica que obteve o pior resultado, gastando um tempo computacional maior do que os demais métodos. O procedimento mais rápido foi o desenvolvido para a regressão por componentes principais, que selecionou 8 componentes.

Tabela 5.3: Erro quadrático médio de predição para cada quantidade de fatores na regressão por componentes principais e na regressão por mínimos quadrados parciais (validações cruzadas)

Componentes	PCR	PLS
1	1.536	1.349
2	1.477	0.786
3	0.301	0.251
4	0.283	0.231
5	0.241	0.229
6	0.241	0.237
7	0.231	0.248
8	0.220	0.275
9	0.225	0.281
10	0.226	0.286

5.1.1 Comentários gerais - Aplicação

Analisando os resultados obtidos para a aplicação, em que o número de covariáveis é maior do que o número de observações, pudemos destacar que a regressão ridge, baseada na proposta de Cule *et al.*, foi a técnica que obteve os melhores resultados preditivos.

Já com relação ao tempo computacional, o método que mais se destacou foi a regressão por componentes principais, que apesar de ter selecionado mais componentes do que a regressão por mínimos quadrados parciais (8 contra 5), demorou apenas 0.406 segundos para modelar e estimar a variável resposta das observações presentes na amostra teste, separada desde o início do processo.

Outra informação interessante, ainda com base no tempo computacional gasto pelos métodos, foi o resultado obtido pelo LASSO, que demorou mais de 2 segundos. Nas simulações realizadas no capítulo anterior, em que o número de covariáveis era menor do

que o número de observações, o LASSO obteve tempo computacional médio menor do que a regressão ridge. Entretanto, para este caso, em que o número de covariáveis é maior do que o número de observações, a regressão ridge foi mais rápida do que o LASSO.

Capítulo 6

Considerações Finais

Com base no trabalho apresentado foi possível conhecer e desenvolver quatro diferentes técnicas de predição adequadas para situações em que há sérios problemas de multicolinearidade e/ou de alta dimensionalidade, incluindo casos em que o número de covariáveis é maior do que o número de observações. Estas situações podem causar resultados instáveis nos métodos de regressão por mínimos quadrados ordinários, além do aumento da variância dos coeficientes estimados.

Com relação aos resultados obtidos, visando apenas questões preditivas, foi possível destacar os seguintes padrões, apresentados na Parte I do Capítulo 3, em que o número de covariáveis era menor do que o de observações:

- Quando há um grande número de pequenos efeitos, a regressão ridge é a melhor opção;
- Quando há um grande número de grandes efeitos, a regressão ridge também é a melhor opção;
- Quando há um pequeno número de grandes efeitos, o LASSO é definitivamente a melhor opção;
- Quando há um moderado número de moderados efeitos, o LASSO também é a melhor opção;
- Já quando há variação na intensidade dos efeitos, ou seja, há pequenos, moderados e grandes efeitos ao mesmo tempo, o LASSO será a melhor opção dependendo da quantidade de efeitos não nulos que houver. Quanto menor for essa quantidade, melhores serão os resultados obtidos pelo LASSO se comparado as outras três técnicas.

Uma possível opção para identificar previamente de que forma são os coeficientes, baseados nas características analisadas, seria ajustar um modelo de regressão por mínimos quadrados ordinários e observar os valores ajustados. Isto permitiria ter uma ideia de como os coeficientes podem ser e ajudar na tomada de decisão de qual técnica utilizar para fins preditivos.

Para os casos de alta dimensionalidade, em que o número de covariáveis era maior do que o número de observações, as simulações foram inconclusivas, ou seja, não foi possível destacar nenhum padrão. Para cada situação, um modelo diferente se destacou mais. Dessa forma, foi desenvolvida uma aplicação com dados reais (não simulados), englobando estes casos. Para esta aplicação a regressão ridge, baseada na proposta de Cule *et al.* (2012), foi a técnica que obteve os melhores resultados preditivos.

Levando em consideração o tempo computacional médio gasto para os modelos serem criados e utilizados para realizar predições, a regressão por componentes principais foi o melhor método, gastando, em média, menos tempo do que as outras três técnicas. A regressão ridge foi a técnica mais demorada nas simulações, em que o número de covariáveis era menor do que o número de observações. Já na aplicação, em que havia maior número de preditores do que de observações, o LASSO foi a técnica mais lenta.

Uma outra característica analisada foi a qualidade que o LASSO tem de selecionar as covariáveis teoricamente significativas. Com base nos resultados obtidos, foi possível destacar que a técnica em questão consegue selecionar os efeitos significativos com grande eficiência. Entretanto, quando houve aumento da proporção de efeitos teoricamente não-nulos, a metodologia apresentou uma tendência a incluir com maior frequência efeitos que não seriam importantes, para os quais os parâmetros de regressão foram fixados como iguais a 0.

Vale ressaltar que apesar dos resultados obtidos, a escolha da técnica a ser utilizada vai muito além de todas essas questões. O pesquisador deve analisar, desde o início, as suas preferências e os seus objetivos. Exemplo disso seria o caso em que o pesquisador tivesse o interesse de realizar uma seleção de covariáveis para obter um modelo mais interpretável. Neste caso, o LASSO se tornaria mais razoável, com base em suas próprias características. Outro exemplo seria ter que trabalhar com mais de uma variável resposta. Nesta situação, o PLS seria a opção mais plausível.

Referências Bibliográficas

[1] Acharjee, A.; Finkers, R.; Visser, R. GF.; Maliepaard, C. *Comparison of Regularized Regression Methods for Omics Data*, Metabolomics, 2013.

[2] Bolfarine, H.; Sandoval, M. C. *Introdução à Inferência Estatística*, 2001

[3] Bussab, W. O.; Morettin, P. A. Estatística Básica, 1987.

[4] Cule, E.; De Iorio, M. A semi-automatic method to guide the choice of ridge parameter in ridge regression, Annals of Applied Statistics, 2012.

[5] De Iorio, M.; Ebbles, T. M. D.; Stephens, D. A. *Statistical Techniques in Metabolomic Profiling*, In Handbook of statistical genetics, 3rd ed. Volume 1. Chapter 11 John Wiley & Sons, 2007.

[6] Duzan, H.; Shariff, N. S. B. M. *Ridge Regression for Solving the Multicollinearity Problem: Review of Methods and Models*, Journal of Applied Sciences 15 (3): 392-404, 2015.

[7] Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. *Least Angle Regression*, The Annals of Statistics, Vol. 32, No. 2, 407-499, 2004.

[8] Dormann, C. F; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J. R. G.; Gruber, B.; Lafourcade, B.; Leitão, P. J.; Münkemüller, T.; McClean, C.; Osborne, P. E.; Reineking, B.; Schröder, B.; Skidmore, A. K.; Zurell, D.; Lautenbach, S. *Collinearity: a review of methods to deal with it and a simulation study evaluating their performance*, Ecography 36: 027-046, 2013.

[9] Geladi, P.; Kowlaski, B. *Partial least square regression: A tutorial*, Analytica Chimica Acta, 35: 117, 1986.

[10] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer, 2009.

[11] Hill, R. C.; Fomby, T. B.; Johnson, S. R. *Component selection norms for principal components regression*, Commun. Statist., A6: 309-334, 1977.

[12] Hoerl, A. E.; Kennard, R. W. *Ridge regression: Applications to nonorthogonal problems*, Technometrics 12: 69-82, 1970b.

[13] Hoerl, A. E.; Kennard, R. W. *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics 12: 55-67, 1970a.

[14] Hoerl, A. E.; Kennard, R. W.; Baldwin, K. F. *Ridge Regression: some simulations*, Communications in Statistics - Theory and Methods 4: 105-123, 1975.

[15] Höskuldsson, A. *PLS regression methods*, Journal of Chemometrics, 2: 211-228, 1988.

[16] Jolliffe, I. T. *Principal Component Analysis*, Second Edition, Springer, 2002.

[17] Mertens, B.; Fearn, T.; Thompson, M. *The efficient crossvalidation of principal components applied to principal component regression*, Statist. Comput., 5: 227-235, 1995.

[18] Mevik, B. H.; Wehrens, R. *The pls Package: Principal Component and Partial Least Squares Regression in R*, Journal of Statistical Software, Volume 18, Issue 2, 2007.

[19] Montgomery, D. C.; Peck, E. A.; Vining, G. G. *Introduction to Linear Regression Analysis*, Third Edition. Hoboken: Wiley, 2001.

[20] R Core Team *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2015.

[21] Rawlings, J. O.; Pantula, S. G.; Dickey, D. A. *Applied Regression Analysis: A Research Tool*, Second Edition, Springer, 1998.

[22] Silva, A. V. L. *Alternativas e comparações de modelos lineares para estimação da biomassa verde de Bambusa vulgaris na existência de multicolinearidade*, 2008.

[23] Tibshirani, R. *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), Volume 58, Issue 1, 267-288, 1996.

[24] Tobias, R. D. *An Introduction to Partial Least Squares Regression*, SAS Institute Inc., Cary, NC.

[25] Wold, S.; Sjostron, M.; Eriksson, L. *PLS-regression: a basic tool of chemometrics*, Chemometrics and Intelligent Laboratory Systems, 58: 109-130, 2001.

[26] Wong, K. Y.; Chiu, S. N. *An iterative approach to minimize the mean squared error in ridge regression*, Comput Stat, 30: 625-639, 2015.