LEANDRO DA SILVA PEREIRA

GEOMETRIA DOS MÉTODOS DE REGRESSÃO LARS, LASSO E ELASTIC NET COM UMA APLICAÇÃO EM SELEÇÃO GENÔMICA

LAVRAS - MG

2017

LEANDRO DA SILVA PEREIRA

GEOMETRIA DOS MÉTODOS DE REGRESSÃO LARS, LASSO E ELASTIC NET COM UMA APLICAÇÃO EM SELEÇÃO GENÔMICA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador
Dr. Lucas Monteiro Chaves

LAVRAS - MG
2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Pereira, Leandro da Silva.
Geometria dos métodos de regressão LARS, LASSO e Elastic Net com uma aplicação em seleção genômica. / Leandro da Silva Pereira. - 2017.
167 p. : il.

Orientador(a): Lucas Monteiro Chaves.

Tese (doutorado) - Universidade Federal de Lavras, 2017.
Bibliografia.

1. Seleção de covariáveis. 2. Geometria de modelos lineares. 3. Regressão Elastic Net. I. Chaves, Lucas Monteiro. . II. Título.

LEANDRO DA SILVA PEREIRA

GEOMETRIA DOS MÉTODOS DE REGRESSÃO LARS, LASSO E ELASTIC NET COM UMA APLICAÇÃO EM SELEÇÃO GENÔMICA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 27 de Janeiro de 2017.

Dr. Lucas Monteiro Chaves (Orientador)	UFLA
Dra. Thelma Sáfadi	UFLA
Dr. Daniel Furtado Ferreira	UFLA
Dr. José Airton Rodrigues Nunes	UFLA
Dr. Fabyano Fonseca e Silva	UFV

LAVRAS - MG
2017

“Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota.”

Santa Madre Teresa de Calcutá

AGRADECIMENTOS

A Deus, pelo sol que ilumina os meus dias, pelos corpos celestes que clareiam as minhas noites mais escuras, pelas chuvas que lavam a minha alma, pelas primaveras que colorem a minha vida com as suas flores, pelos pássaros que alegram o meu espírito com os seus cantos, pelos alimentos que fortificam o meu corpo, pela verdade que aos poucos me liberta, pelo trabalho que me faz sentir útil, pela vida que me proporciona novas experiências e oportunidades de crescimento, e por tudo que ainda não tenho consciência de que ELE faz por mim.

Aos meus pais Messias e Valdelice que, pelo exemplo de humildade e dignidade, me mostram o sentido da vida.

À minha eterna companheira Mayumi, sinônimo de vida e esperança, por ser meu forte nas horas difíceis e meu encanto nas horas de paz.

À minha irmã Celiani e sua família, pelo apoio direto e indireto.

A todos os meus familiares e amigos, pelo apoio e carinho.

Ao meu orientador Lucas Monteiro Chaves, que pelos conhecimentos e esclarecimentos intelectuais e morais confiados a mim, me ensinou o verdadeiro sentido de se ensinar.

Ao Prof. Devanil Jaques de Souza, que de forma síncrona com as ideias do professor Lucas, contribui para minha formação moral e intelectual.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística (DES), pela oportunidade de cursar o doutorado.

Aos professores do programa de pós-graduação em estatística e experimentação agropecuária da UFLA, pelas contribuições na minha formação durante as suas disciplinas, em especial aos professores Renato Ribeiro de Lima, Júlio de Sousa Bueno Filho, Thelma Sáfadi e Daniel Furtado Ferreira.

Às funcionárias do DEX: Kelly, Josi, Maria e Lu, pela amizade e cari-

nho, em especial a Nádia, pelo companheirismo e compreensão nas minhas várias ligações e solicitações burocráticas.

À FAPEMIG, pela concessão da bolsa de estudos no início do doutorado, tornando financeiramente possível a realização do mestrado. Aos colegas da Engenharia de Controle e Automação pelos momentos de lazer e pelos conselhos “etilizados".

A todos os colegas de turma pela companhia nos estudos, nos momentos de lazer e nos momentos de tensão de véspera das provas.

A UTFPR pelo apoio com a licença para finalização do doutorado.

A todos que de alguma forma contribuíram, mesmo inconsciente, para a realização deste trabalho.

RESUMO

Os métodos de estimação e seleção de variáveis em modelos lineares, LASSO (Least Absolute Shrinkage and Selection Operator), LARS (Least Angle Regression) e Elastic Net, são abordados utilizando ênfase em aspectos geométricos. O texto se propõe a ser uma leitura auxiliar aos artigos clássicos de Tibshirani, Hastie, Efron, Zou e Johnstone, apresentando de forma mais detalhada alguns dos resultados citados nos referidos artigos. Tal ponto de vista não ocorre na literatura básica relativa a estes métodos, e neste sentido o trabalho representa uma contribuição original ao assunto. Simulações utilizando a linguagem R (pacote glmnet) são desenvolvidas para se estudar o comportamento dos estimadores. Para um conjunto de dados de suínos ***Sus scrofa***, são analisados 237 marcadores genéticos(SNPs) pelo método Elastic Net, para a resposta relativa ao pH da carne e peso de carcaça.

Palavras-chave: Seleção de covariáveis. Geometria de modelos lineares. Marcadores genéticos. Algoritmo LARS.

ABSTRACT

The methods of estimation and variable selection in linear models, LASSO (Least Absolute Shrinkage and Selection Operator), LARS (Least Angle Regression) and Elastic Net, are addressed using emphasis in terms of a geometric approach. The present work proposes to be an auxiliary reading to the classic articles of Tibshirani, Hastie, Efron, Zou and Johnstone, presenting in more detail some of the results cited in such articles. Such a point of view does not occur in the basic literature regarding these methods, and in this sense the work represents an original contribution to the subject. Simulations using R code (glmnet package) are developed to study the behavior of the estimators. For a data set of ***Sus scrofa*** pork, 237 genetic markers (SNPs) are analyzed using the Elastic Net method, using as response the pork pH and carcass weight.

Keywords: Covariates selection. Linear models geometry. Genetic markers. LARS Algorithm.

Dr. Lucas Monteiro Chaves
Orientador

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com o advento de um número cada vez maior de dados é colocado o problema de se obter novos e mais eficientes métodos estatísticos. Em particular, para a teoria da regressão linear o número elevado de covariáveis demanda, além de um robusto método de estimação dos parâmetros, também um método de seleção de covariáveis. O método de estimação utilizando quadrados mínimos apresenta problemas quando se tem, por exemplo, quasi-colinearidade, isto é, presença de covariáveis altamente correlacionadas. Os dois métodos clássicos mais utilizados para se sanar tal deficiência são a estimação Ridge e o método denominado Subset Selection. No entanto, ambos os métodos apresentam problemas. O método Subset Selection é excessivamente variável, pois é um processo discreto. A regressão Ridge é muito mais estável, mas raramente gera estimativas nulas para os parâmetros. Em razão destes fatos, foi proposto Tibshirani (1996) um novo método denominado LASSO (Least Absolute Shrinkage and Selection Operator). A principal característica deste método é que no processo de encolhimento, a estimativa de muitas covariáveis se anulam, e portanto o método é também um método automático de seleção de covariáveis. Definida em termos de quadrados mínimos penalizados, apresenta um desenvolvimento analítico bastante complexo. Um dos objetivos deste trabalho é apresentar a teoria do método de forma mais geométrica possível. Além disso, várias detalhadas demonstrações, não apresentadas na literatura, são desenvolvidas no sentido de tornar o texto uma referência para os artigos básicos da teoria.

O método LASSO também apresenta problemas, por exemplo, no caso em que se tem mais covariáveis do que observações, o método seleciona no máximo o número de covariáveis igual ao número de observações. Tem-se também que se duas covariáveis são altamente correlacionadas, o método LASSO tem tendência

a selecionar apenas uma delas. No sentido de se obter um método que possua as propriedades do LASSO, mas que consiga minimizar suas deficiências, um novo método denominado Elastic Net é proposto (ZOU; HASTIE, 2005). É também um método baseado na penalização da soma de quadrados. Esta penalização é obtida como uma combinação ponderada entre as penalizações do LASSO e do método Ridge. Este novo método, além das propriedades de seleção de covariáveis, tem a vantagem de fazer seleção por grupo, isto é, possui a tendência de que covariáveis altamente correlacionadas serem simultaneamente selecionadas. Ambos os métodos Lasso e Elastic Net não possuem forma explícita para se determinar suas estimativas. Estas são obtidas através de procedimentos algorítmicos. Nesse sentido foi feito também um estudo bastante aprofundado do algoritmo LARS (Least Angle Regressio) proposto por Efron et al. (2004). Este algoritmo de seleção de modelos é uma evolução dos algoritmos clássicos Forward Stepwise e Forward Stagewise. Possui importância em si mesmo e com pequenas alterações, calcula tanto a estimativa LASSO quanto a estimativa Elastic Net, sendo a base fundamental do pacote *glmnet* na linguagem R.

Sobre a notação utilizada, matrizes serão denotadas por letras maiúsculas $(\mathrm{A}, \mathrm{B}, \mathrm{X})$, vetores coluna por letras minúsculas em negrito $(\mathbf{a}, \mathbf{b}, \mathbf{v}, \boldsymbol{\mu})$, escalares por letras minúsculas em itálico (a, b, γ) e a transposta de uma matriz representada por um apóstrofo junto a esta, isto é, $\mathrm{X}^{\mathrm{t}} = \mathrm{X}'$. Será utilizada também a notação $\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_n)'$ para representação de vetores aleatórios e $\mathbf{y} = (\mathrm{y}_1, ..., \mathrm{y}_\mathrm{n})'$ para valores observados do vetor aleatório $\mathbf{Y}$. Para ser coerente com a notação de alguns artigos, o vetor $\mathbf{y}$ poderá significar também um vetor aleatório. A imagem de uma transformação linear X será representada por $\mathrm{Im(X)}$, a dimensão do subespaço gerado por uma transformação linear X é denotado por $\mathrm{Dim\,(X)}$, o núcleo (ou *Kernel*) de uma transformação é denotado por $\mathrm{Ker\,(X)}$ e o traço de uma matriz

quadrada X é representado por $\mathrm{tr}\,(\mathrm{X})$.

Geometricamente, um vetor linha ou coluna com p elementos pode ser associado com um ponto num espaço p-dimensional. Os elementos no vetor são as coordenadas do ponto. Nesse sentido, um conjunto de n observações pode ser geometricamente interpretado como um vetor $\mathbf{y} = (\mathrm{y}_1, ..., \mathrm{y}_\mathrm{n})'$, de dimensão $\mathrm{n}\times 1$, em que cada coordenada deste corresponde a uma observação. A Figura 1 ilustra este tipo de representação.

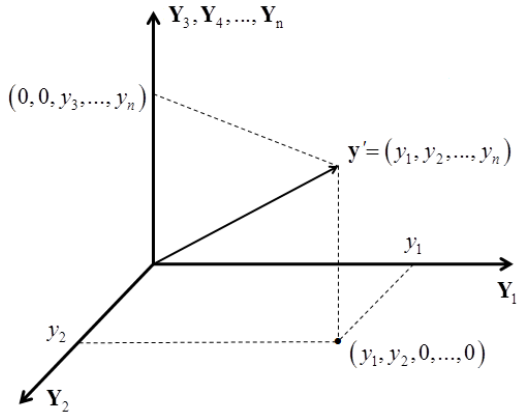


Figura 1 Representação geométrica de um vetor n-dimensional.

Com esta linguagem geométrica, o presente trabalho pretende apresentar em detalhes a teoria dos estimadores LASSO e Elastic Net. Algumas simulações são desenvolvidas para se estudar o comportamento destes estimadores na presença de vários níveis de correlação entre as covariáveis. Uma aplicação em seleção genômica é desenvolvida e analisada.

2 Uma abordagem geométrica à regressão linear múltipla

A regressão linear múltipla admite uma abordagem geométrica que permite um tratamento similar para três teorias importantes: a regressão *ridge*, o LASSO e o *Elasticnet*. Tal abordagem não é usual, sendo que cada uma das teorias anteriormente citadas aparentam demandar resultados teóricos diferentes.

Primeiramente será analisado o caso de regressão linear simples.

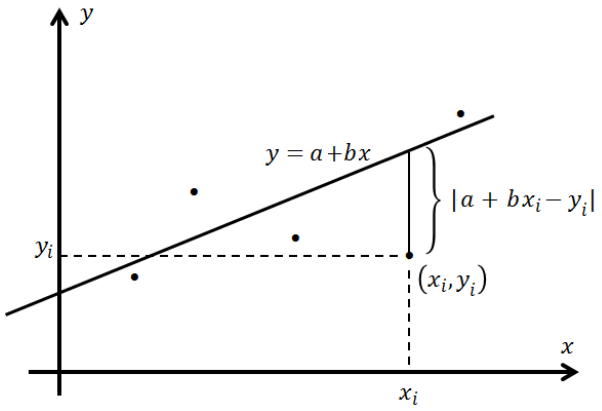


Figura 2 Reta ajustada por quadrados mínimos.

Sejam $(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), \ldots, (\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n})$ os valores observados. Deseja-se então ajustar uma reta $\hat{\mathrm{y}} = \mathrm{bx} + \mathrm{a}$ de tal forma que a equação (2.1) seja minimizada.

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \hat{\mathrm{y}}_i)^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - (a + b\,\mathrm{x_i}))^2 \tag{2.1}$$

A Figura 2 ilustra tal contexto.

O método usual para se obter a e b é derivar (2.1) em relação a estes, igualar as equações derivadas a 0 e resolver o sistema. Uma abordagem geométrica é obtida da seguinte forma:

Se $\mathbf{y}_{\mathrm{n}\times 1} = (\mathrm{y}_1, ..., \mathrm{y}_\mathrm{n})'$, $\boldsymbol{\beta}_{2\times 1} = (a, b)'$, $\boldsymbol{\varepsilon}_{n\times 1} = (\varepsilon_1, ..., \varepsilon_n)'$ e

$$\mathrm{X}_{\mathrm{n}\times 2} = \begin{bmatrix} 1 & \mathrm{x}_1 \\ 1 & \mathrm{x}_2 \\ \vdots & \vdots \\ 1 & \mathrm{x}_\mathrm{n} \end{bmatrix}, \text{ então}$$

$$\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Rightarrow \begin{bmatrix} \mathrm{y}_1 \\ \vdots \\ \mathrm{y}_\mathrm{n} \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{x}_1 \\ \vdots & \vdots \\ 1 & \mathrm{x}_\mathrm{n} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_\mathrm{n} \end{bmatrix}$$

isto é, $\boxed{\mathrm{y_i} = a + b\mathrm{x_i} + \varepsilon_\mathrm{i}}$, $\mathrm{i} = 1, ..., \mathrm{n}$. Portanto, reescrevendo (2.1) tem-se

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - (a + b\mathrm{x_i}))^2 = \left\|\mathbf{y}_{\mathrm{n}\times 1} - \mathrm{X}_{\mathrm{n}\times 2}\boldsymbol{\beta}_{2\times 1}\right\|^2.$$

Em termos de transformações lineares em que a matriz $\mathrm{X}_{\mathrm{n}\times 2}$ atua como uma transformação do espaço de parâmetros para o espaço dos dados, o conceito pode ser representado pela Figura 3.

Assim,

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - (a + b\mathrm{x_i}))^2 = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2.$$

Logo, minimizar $\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - (a + b\mathrm{x_i}))^2$ equivale a minimizar $\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$ que, claramente, ocorre quando $\mathrm{X}\boldsymbol{\beta}$ é a projeção ortogonal do vetor de dados $\mathbf{y}$ no subespaço $\mathrm{Im(X)}$. Este vetor projeção será denominado $\mathrm{P}_{\mathrm{Im(X)}}\mathbf{y}$, também denotado por $\hat{\mathbf{y}}$, vetor de dados ajustados. Note que, como o posto (coluna) de X é 2, a dimensão da imagem de X é $\mathrm{Dim(Im(X))} = 2$.

Na regressão linear múltipla, se tem várias covariáveis, isto é, $(\mathrm{x}_1, ..., \mathrm{x}_\mathrm{p})$, que foram observadas n vezes. Deseja-se então ajustar o modelo

$$\boxed{\mathrm{y_i} = \beta_0 + \beta_1\mathrm{x_{i1}} + \beta_2\mathrm{x_{i2}} + \cdots + \beta_\mathrm{p}\mathrm{x_{ip}} + \varepsilon_\mathrm{i}} \quad \mathrm{i} = 1, ..., \mathrm{n}$$

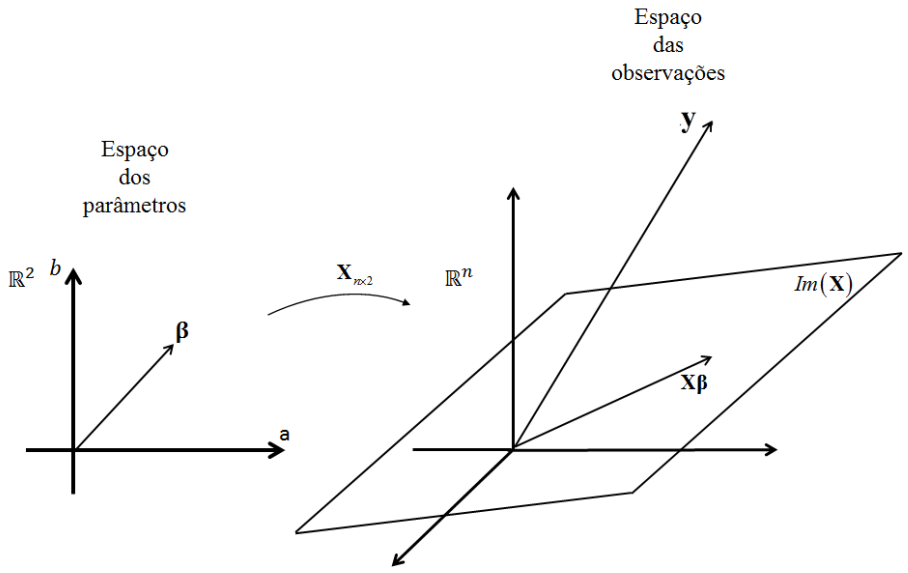


Figura 3 Geometria da regressão linear simples.

o que matricialmente equivale a

$$\begin{pmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_\mathrm{n} \end{pmatrix}_{\mathrm{n}\times 1} = \begin{pmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots & \mathrm{x}_{1\mathrm{p}} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2\mathrm{p}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}1} & \mathrm{x}_{\mathrm{n}2} & \cdots & \mathrm{x}_{\mathrm{np}} \end{pmatrix}_{\mathrm{n}\times\mathrm{p}+1} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_\mathrm{p} \end{pmatrix}_{\mathrm{p}+1\times 1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_\mathrm{n} \end{pmatrix}_{\mathrm{n}\times 1} \tag{2.2}$$

e se quer minimizar $\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y}_\mathrm{i} - \beta_0 - \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{x}_{\mathrm{ij}}\beta_\mathrm{j} \right)^2 = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$. Observe então que, geometricamente, o modelo para o vetor de médias do vetor $\mathbf{y}$ é simplesmente a definição do subespaço vetorial $\mathrm{Im(X)}$. O valor mínimo é obtido pela projeção ortogonal do vetor de dados $\mathbf{y}$ no subespaço $\mathrm{Im(X)}$, que possui nesse caso dimensão $\mathrm{p}+1$. A Figura 4 ilustra esta situação.

Geometricamente a representação é similar a Figura 3, diferindo apenas

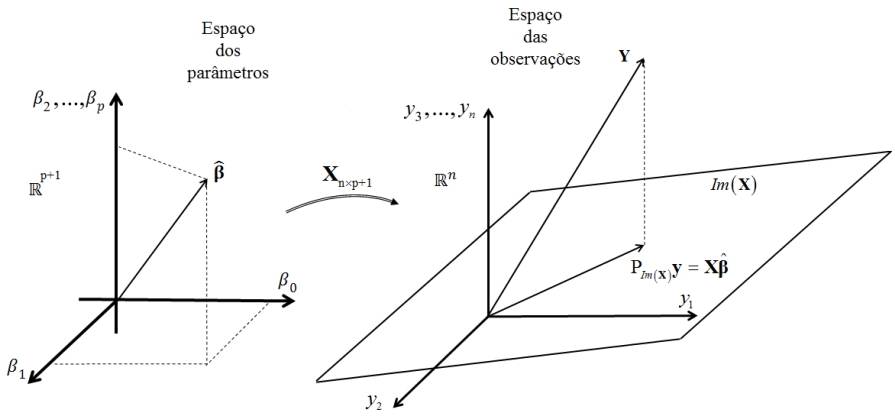


Figura 4 Geometria da regressão linear múltipla.

na dimensão do espaço dos parâmetros que agora passa a ser o $\mathbb{R}^{\mathrm{p}}$ (ver Figura 4). Considerando o posto de X igual a p (posto completo), tem-se que $\mathrm{Dim}\,(\mathrm{Im}\,(\mathrm{X})) = \mathrm{p}$. Para se obter uma expressão para $\mathrm{P}_{\mathrm{Im(X)}}\mathbf{y}$ é necessário a teoria das matrizes de projeção.

Definição 1. *Uma matriz quadrada é dita matriz de projeção ou projetor se é idempotente, isto é,* $\mathrm{A}^2 = \mathrm{A}$.

Pela definição 1, tem-se que um projetor A restrito a $\mathrm{Im}\,(\mathrm{A})$ é a identidade, ou seja,

$$\begin{aligned} \mathrm{A}\,(\mathrm{A}\mathbf{z}) &= \mathrm{A}^2\mathbf{z} \\ &= \mathrm{A}\mathrm{z}. \end{aligned}$$

Observe que, se I é a matriz identidade, $\mathrm{I} - \mathrm{A}$ também é um projetor pois

$$\begin{aligned} (\mathrm{I}-\mathrm{A})^2 &= \mathrm{I} - 2\mathrm{A} + \mathrm{A}^2 \\ &= \mathrm{I} - 2\mathrm{A} + \mathrm{A} \\ &= \mathrm{I} - \mathrm{A}. \end{aligned}$$

Como

$$\begin{aligned} \mathrm{A}\left(\left(\mathrm{I}-\mathrm{A}\right)\mathbf{z}\right) &= \mathrm{A}\left(\mathbf{z}-\mathrm{A}\mathbf{z}\right) \\ &= \mathrm{A}\mathbf{z}-\mathrm{A}^2\mathbf{z} \\ &= \mathrm{A}\mathbf{z}-\mathrm{A}\mathbf{z} \\ &= \mathbf{0}, \end{aligned}$$

segue que $\mathrm{Im}\left(\mathrm{I}-\mathrm{A}\right) \subset \mathrm{Ker}\left(\mathrm{A}\right)$. Além do mais, se $\mathbf{z} \in \mathrm{Ker}\left(\mathrm{A}\right)$ então

$$\begin{aligned} \left(\mathrm{I}-\mathrm{A}\right)\mathbf{z} &= \mathbf{z}-\mathrm{A}\mathbf{z} \\ &= \mathbf{z}-\mathbf{0}. \end{aligned}$$

Logo, $\mathbf{z} \in \mathrm{Im}\left(\mathrm{I}-\mathrm{A}\right)$ e, portanto, $\mathrm{Im}\left(\mathrm{I}-\mathrm{A}\right) = \mathrm{Ker}\left(\mathrm{A}\right)$. Assim, segue então que $\mathrm{Im}\left(\mathrm{A}\right) = \mathrm{Ker}\left(\mathrm{I}-\mathrm{A}\right)$.

Definição 2. *Uma matriz de projeção* A *é dita um projetor ortogonal se, para um dado vetor* $\mathbf{w}$*,* $\mathrm{A}\mathbf{w}-\mathbf{w}$ *é perpendicular ao subespaço* $\mathrm{Im}\left(\mathrm{A}\right)$.

Proposição 1. *Uma matriz de projeção é simétrica se, e somente se, é um projetor ortogonal.*

Demonstração. Suponha A simétrica e $\mathbf{v}$ e $\mathbf{w}$ vetores quaisquer. Utilizando as propriedades de produto interno e sendo $\mathrm{A}^2 = \mathrm{A}$, tem-se que

$$\begin{aligned} \left\langle \mathbf{v}-\mathrm{A}\mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle &= \left\langle \mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle - \left\langle \mathrm{A}\mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle \\ &= \left\langle \mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle - \left\langle \mathbf{v}, \mathrm{A}'\mathrm{A}\mathbf{w}\right\rangle \\ &= \left\langle \mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle - \left\langle \mathbf{v}, \mathrm{A}\mathrm{A}\mathbf{w}\right\rangle \\ &= \left\langle \mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle - \left\langle \mathbf{v}, \mathrm{A}^2\mathbf{w}\right\rangle \\ &= \left\langle \mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle - \left\langle \mathbf{v}, \mathrm{A}\mathbf{w}\right\rangle \end{aligned}$$

$$= \mathbf{0},\ \forall \mathbf{v}, \mathbf{w}.$$

Portanto, $\mathbf{v} - \mathrm{A}\mathbf{v}$ é perpendicular a $\mathrm{Im}\,(\mathrm{A})$. Seja agora $\mathrm{A}\mathbf{v} - \mathbf{v}$ perpendicular a $\mathrm{Im}\,(\mathrm{A})$. Tem-se então que, para $\mathbf{w}$ qualquer,

$$\begin{aligned} 0 &= \langle \mathrm{A}\mathbf{v} - \mathbf{v}, \mathrm{A}\mathbf{w} \rangle \\ &= \langle \mathrm{A}\mathbf{v}, \mathrm{A}\mathbf{w} \rangle - \langle \mathbf{v}, \mathrm{A}\mathbf{w} \rangle , \end{aligned}$$

ou seja,

$$\langle \mathrm{A}\mathbf{v}, \mathrm{A}\mathbf{w} \rangle = \langle \mathbf{v}, \mathrm{A}\mathbf{w} \rangle . \tag{2.3}$$

Também, para $(\mathrm{A}\mathbf{w}) - \mathbf{w}$ perpendicular a $\mathrm{A}\mathbf{v}$,

$$\begin{aligned} 0 &= \langle \mathrm{A}\mathbf{w} - \mathbf{w}, \mathrm{A}\mathbf{v} \rangle \\ &= \langle \mathrm{A}\mathbf{w}, \mathrm{A}\mathbf{v} \rangle - \langle \mathbf{w}, \mathrm{A}\mathbf{v} \rangle , \end{aligned}$$

ou seja, $\langle \mathrm{A}\mathbf{w}, \mathrm{A}\mathbf{v} \rangle = \langle \mathbf{w}, \mathrm{A}\mathbf{v} \rangle$ e portanto

$$\langle \mathrm{A}\mathbf{v}, \mathrm{A}\mathbf{w} \rangle = \langle \mathrm{A}\mathbf{v}, \mathbf{w} \rangle . \tag{2.4}$$

Consequentemente, das equações (2.3) e (2.4)

$$\langle \mathrm{A}\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathrm{A}\mathbf{w} \rangle .$$

□

Uma interpretação geométrica relativa à matriz transposta A' (operador adjunto) pode ser visualizada na Figura 5.

Para se obter uma expressão para a projeção $\mathrm{P}_{\mathrm{Im(X)}}\mathbf{y}$, será apresentada na Proposição 2 uma dedução baseada apenas em argumentos geométricos.

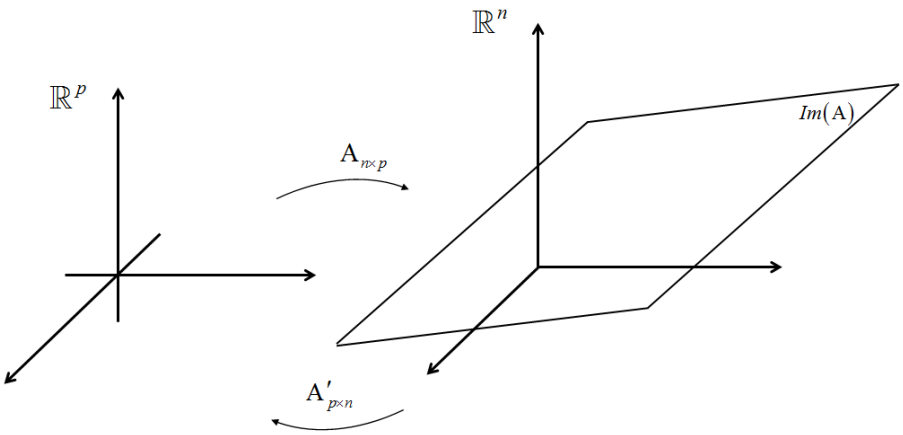


Figura 5 Interpretação geométrica das matrizes A e A′

Proposição 2. *A projeção ortogonal do vetor de dados* $\mathbf{y}$ *no subespaço* $\mathrm{Im}\,(\mathrm{X})$ *é*

$$\mathrm{P}_{\mathrm{Im}(\mathrm{X})}\mathbf{y} = \mathrm{X}\big(\mathrm{X}'\mathrm{X}\big)^{-1}\mathrm{X}'\mathbf{y}.$$

Demonstração. Seja $\mathbf{y} = (\mathrm{y}_1, ..., \mathrm{y}_\mathrm{n})$. Em razão dos erros, tem-se que a probabilidade $\mathrm{P}\left[\mathbf{y} \in \mathrm{Im}\,(\mathrm{X})\right] = 0$. Como queremos minimizar $\mathrm{L}\,(\boldsymbol{\beta}) = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$ e o vetor $\mathrm{X}\boldsymbol{\beta}$ pertence à imagem de X, segue que $\mathrm{L}\,(\boldsymbol{\beta})$ é mínimo quando $\mathrm{X}\boldsymbol{\beta}$ é a projeção ortogonal de $\mathbf{y}$ na $\mathrm{Im}\,(\mathrm{X})$, com notação $\mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})$, e, consequentemente, existe um vetor $\hat{\boldsymbol{\beta}}$ de parâmetros tal que $\mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})$ já que a matriz X é de posto completo. Considere ainda o vetor de parâmetros $\tilde{\boldsymbol{\beta}} = \mathrm{X}'\mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})$. Observe que, como $\mathbf{y} - \mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})$ é ortogonal a qualquer vetor na imagem de X, $\mathbf{y} - \mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})$ está no espaço nulo de X, isto é, $\mathrm{Ker}\,(\mathrm{X})$. Segue então que $\mathrm{X}'\left(\mathbf{y} - \mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})\right) = \mathbf{0}$. Portanto, pode-se escrever que

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \mathrm{X}'\mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y}) \\
&= \mathrm{X}'\mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y}) + \mathrm{X}'\left(\mathbf{y} - \mathrm{P}_{\mathrm{Im}(\mathrm{X})}\,(\mathbf{y})\right) \\
&= \mathrm{X}'\mathbf{y}. \qquad (2.5)
\end{aligned}$$

Como $\mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{P}_{\mathrm{Im(X)}}\,(\mathbf{y})$, pode-se também escrever que

$$\boldsymbol{\beta} = \mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}}. \tag{2.6}$$

Assim, de (2.7) e (2.6) segue que

$$\begin{aligned} \boldsymbol{\beta} &= \mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}} \\ &= \mathrm{X}'\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= \big(\mathrm{X}'\mathrm{X}\big)^{-\mathbf{1}}\mathrm{X}'\mathbf{y}, \end{aligned}$$

conforme representado na Figura 4. □

Segue da Proposição 2 que a estimativa do vetor de parâmetros $\boldsymbol{\beta}$, que nos dá o ajuste de quadrados mínimos, é

$$\hat{\boldsymbol{\beta}} = \big(\mathrm{X}'\mathrm{X}\big)^{-1}\mathrm{X}'\mathbf{y},$$

pois

$$\begin{aligned} \mathrm{X}\hat{\boldsymbol{\beta}} &= \mathrm{X}\big(\mathrm{X}'\mathrm{X}\big)^{-1}\mathrm{X}'\mathbf{y} \\ &= \mathrm{P}_{\mathrm{Im(X)}}\mathbf{y}. \end{aligned}$$

O modelo ajustado fica então da forma

$$\boxed{\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + \hat{\beta}_2\mathbf{x}_2 + \cdots + \hat{\beta}_{\mathrm{p}}\mathbf{x}_{\mathrm{p}}}.$$

2.1 O triângulo fundamental

A geometria dos modelos de regressão linear múltipla é definida por um triângulo retângulo conforme destacado na Figura 6, em que SST, SSR e SSE são, respectivamente, abreviações do inglês para *Total Sum of Square*, *Regression Sum of Square* e *Explained Sum of Square*, conforme Rencher (2008). Portanto, segue

que

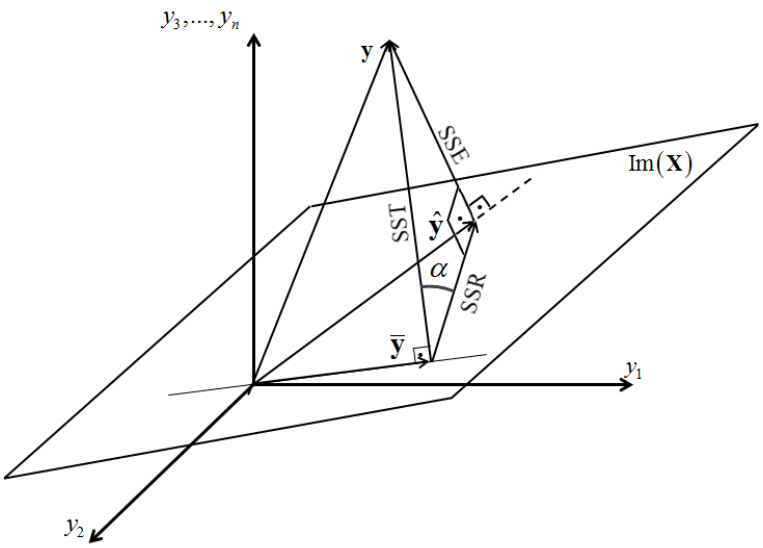


Figura 6 O triângulo fundamental.

$$(\mathbf{y}-\hat{\mathbf{y}})+(\hat{\mathbf{y}}-\bar{\mathbf{y}})=\mathbf{y}-\bar{\mathbf{y}}$$

$$(\mathbf{y}-\hat{\mathbf{y}})'(\hat{\mathbf{y}}-\bar{\mathbf{y}})=0$$

$$\|\mathbf{y}-\hat{\mathbf{y}}\|^2+\|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|^2=\|\mathbf{y}-\bar{\mathbf{y}}\|^2,$$

e então

$$SST=\|\mathbf{y}-\bar{\mathbf{y}}\|^2=\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}})^2=(\text{hipotenusa})^2$$

$$SSR=\|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|^2=\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i}-\bar{\mathrm{y}})^2=(\text{cateto adjacente})^2$$

$$SSE=\|\mathbf{y}-\hat{\mathbf{y}}\|^2=\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^2=(\text{cateto oposto})^2.$$

Logo,

$$\mathrm{SST}\ =\ \mathrm{SSR}\ +\ \mathrm{SSE}.$$

As somas de quadrados podem ser descritas em termos das projeções em seus respectivos subespaços, conforme se segue:

$$\begin{aligned}
\mathrm{SSR} &= \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 \\
&= (\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \\
&= \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} - \frac{1}{\mathrm{n}}\mathrm{J}\mathbf{y}\right)' \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} - \frac{1}{\mathrm{n}}\mathrm{J}\mathbf{y}\right) \\
&= \mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathrm{J}\right)' \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathrm{J}\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathrm{J}\right)^2\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathrm{J} - \frac{1}{\mathrm{n}}\mathrm{J}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' + \left(\frac{1}{\mathrm{n}}\mathrm{J}\right)^2\right)\mathbf{y}.
\end{aligned}$$

Assume-se aqui que $\overrightarrow{\mathbf{1}} \in \mathrm{Im}(\mathrm{X})$. Esta é a hipótese do modelo linear com intercepto, a qual é mais utilizada nos livros texto. Neste caso, como $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ é a projeção em $\mathrm{Im}(\mathrm{X})$ e $\frac{1}{\mathrm{n}}\mathrm{J}$ a projeção no subespaço gerado por $\overrightarrow{\mathbf{1}}$, então $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathrm{J}\mathbf{y} = \frac{1}{\mathrm{n}}\mathrm{J}\mathbf{y}$ para todo $\mathbf{y}$, i.e. , $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathrm{J} = \frac{1}{\mathrm{n}}\mathrm{J}$. Em contra partida, $\frac{1}{\mathrm{n}}\mathrm{J}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} = \frac{1}{\mathrm{n}}\mathrm{J}\hat{\mathbf{y}} = \overline{\hat{\mathbf{y}}}$. Tem-se então que

Proposição 3. *Se* $\overrightarrow{\mathbf{1}} = (1, 1, \cdots, 1)' \in \mathrm{Im}(\mathrm{X})$ *então a média do vetor ajustado é igual à média do vetor observado, i.e.,* $\overline{\hat{\mathrm{y}}} = \bar{\mathrm{y}}$.

Demonstração. Considere a seguinte decomposição em soma direta ortogonal (Figura 7). Se $\mathrm{V}_{\overrightarrow{\mathbf{1}}}$ é o subespaço gerado pelo vetor $\overrightarrow{\mathbf{1}}$, $\mathrm{V}_{\overrightarrow{\mathbf{1}}} = \{(\mathrm{a}, \mathrm{a}, \cdots, \mathrm{a}),\ \mathrm{a} \in \mathbb{R}\}$, então

$$\mathbb{R}^{\mathrm{n}} = \mathrm{V}_{\overrightarrow{\mathbf{1}}} \oplus \left\{\mathrm{V}_{\overrightarrow{\mathbf{1}}}{}^{\perp} \cap \mathrm{Im}\,(\mathrm{X})\right\} \oplus (\mathrm{Im}\,(\mathrm{X}))^{\perp}.$$

Assim, tem-se que $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3$ com $\mathbf{v}_1$, $\mathbf{v}_2$ e $\mathbf{v}_3$ pertencendo respectivamente a $\mathrm{V}_{\overrightarrow{\mathbf{1}}}$, $\left(\mathrm{V}_{\overrightarrow{\mathbf{1}}}{}^{\perp} \cap \mathrm{Im}\,(\mathrm{X})\right)$ e $\mathrm{Im}(\mathrm{X})^{\perp}$. Também, $\mathrm{P}_{\mathrm{V}_1}(\mathbf{v}) = \mathbf{v}_1$ e

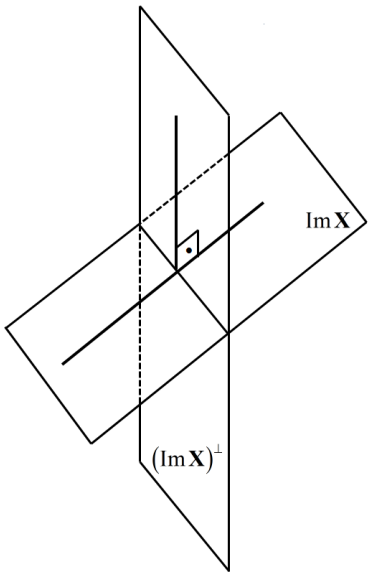


Figura 7 Projeções em termos de somas direta.

$\mathrm{P}_{\mathrm{Im(X)}}(\mathbf{v}) = \mathbf{v}_1 + \mathbf{v}_2$. Portanto,

$$\mathrm{P}_{\mathrm{V}_1}\left(\mathrm{P}_{\mathrm{Im(X)}}(\mathbf{v})\right) = \mathrm{P}_{\mathrm{V}_1}(\mathbf{v}).$$

Se $\mathbf{y}$ é o vetor observado, $\mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y}) = \hat{\mathbf{y}}$ e $\mathrm{P}_{\mathrm{V}_1}(\mathbf{y}) = \bar{\mathbf{y}}$. Desta forma,

$$\bar{\mathbf{y}} = \mathrm{P}_{\mathrm{V}_1}(\mathbf{y}) = \mathrm{P}_{\mathrm{V}_1}\left(\mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y})\right) = \mathrm{P}_{\mathrm{V}_1}\left(\hat{\mathbf{y}}\right) = \hat{\bar{\mathbf{y}}}.$$

□

Uma outra demonstração mais geométrica pode ser baseada nos dois triângulos representados na Figura 8, conforme se segue.

Suponha que a projeção ortogonal de $\hat{\mathbf{y}}$ na direção de $\overrightarrow{\mathbf{1}}$ não seja $\bar{\mathbf{y}}$. Tem-se então os dois triângulos retângulos $\mathrm{c}^2 = \mathrm{b}^2 + \mathrm{a}^2$ e $\mathrm{e}^2{=}\mathrm{d}^2 + \mathrm{a}^2$. Como $\mathrm{c} < \mathrm{e}$ e $\mathrm{d} < \mathrm{b}$ uma vez que as projeções são ortogonais, tem-se $\mathrm{c}^2 < \mathrm{e}^2$ e $\mathrm{d}^2 < \mathrm{b}^2$. Então, $\mathrm{b}^2 + \mathrm{a}^2 < \mathrm{e}^2{=}\mathrm{a}^2 + \mathrm{d}^2 \Rightarrow \mathrm{b}^2 < \mathrm{d}^2$, o que é uma contradição.

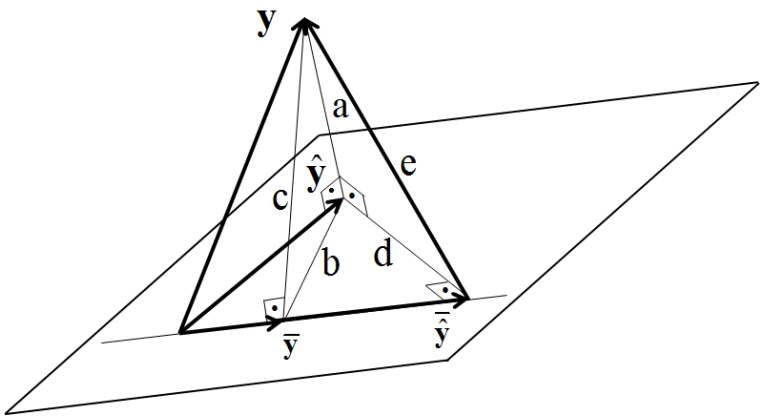


Figura 8 Demonstração geométrica da Proposição 3

Segue agora um exemplo em que $\bar{\mathbf{y}} \neq \overline{\hat{\mathbf{y}}}$.

Para $\mathrm{X} = \begin{pmatrix} 1 & 1 \\ 1 & -2 \\ -2 & 1 \end{pmatrix}$, $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, se tem $\overrightarrow{\mathbf{1}} \notin \mathrm{Im}\,(\mathrm{X})$. $\hat{\mathbf{y}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$, e então $\overline{\hat{\mathrm{y}}} = 0 \neq 2 = \bar{\mathrm{y}}$.

Pela Proposição 3, $\frac{1}{\mathrm{n}}\mathrm{JX}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} = \frac{1}{\mathrm{n}}\mathrm{J}\hat{\mathbf{y}} = \overline{\hat{\mathbf{y}}} = \bar{\mathbf{y}} = \frac{1}{\mathrm{n}}\mathrm{J}\mathbf{y}$.

Portanto, a SSR resulta em

$$\mathrm{SSR} = \mathbf{y}' \left(\mathrm{X}\big(\mathrm{X}'\mathrm{X}\big)^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathrm{J} \right) \mathbf{y}.$$

Para aqueles que preferem uma abordagem mais algébrica, um bom desafio é provar algebricamente que $\frac{1}{\mathrm{n}}\mathrm{JX}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathrm{J} = \frac{1}{\mathrm{n}}\mathrm{J}$. A SSE é calculada então por

$$\begin{aligned} \mathrm{SSE} &= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ &= (\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}}) \end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbf{y} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathbf{y}\right)' \left(\mathbf{y} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathbf{y}\right) \\
&= \left(\left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)\mathbf{y}\right)' \left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)' \left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)^{2}\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)\mathbf{y}.
\end{aligned}
$$

Novamente, a identidade fundamental pode ser demonstrada:

$$
\begin{aligned}
\mathrm{SSE} + \mathrm{SSR} &= \mathbf{y}'\left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right)\mathbf{y} + \mathbf{y}'\left(\mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathrm{J}\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' + \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathrm{J}\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{I} - \frac{1}{\mathrm{n}}\mathrm{J}\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathbf{y} - \bar{\mathbf{y}}\right) \\
&= \left(\mathbf{y}' - \bar{\mathbf{y}}\right)'\left(\mathbf{y} - \bar{\mathbf{y}}\right) \\
&= \mathrm{SST}.
\end{aligned}
$$

2.2 Reparametrização em sistemas lineares

Em problemas de regressão linear, é usual que por exemplo se altere a unidade de medida de covariáveis. Neste caso é necessário uma reparametrização do modelo. Formalmente, a reparametrização pode ser descrita como se segue. Considerando o modelo $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ em que $\mathrm{E}\left[\boldsymbol{\varepsilon}\right] = 0$ e $\mathrm{cov}\left[\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\right] = \sigma^2\mathrm{I}$, o estimador de quadrados mínimos é dado por $\hat{\boldsymbol{\beta}} = \left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathbf{y}$. Uma reparametrização linear é aquela em que os novos parâmetros são funções lineares dos parâmetros originais. Suponha que os componentes do vetor $\boldsymbol{\beta}$ sejam combinações lineares dos componentes de um vetor $\boldsymbol{\gamma}_{\mathrm{p}\times 1}$, expressas por $\boldsymbol{\beta} = \mathrm{A}_{\mathrm{p}\times\mathrm{p}}\boldsymbol{\gamma}$. Desta forma,

$$
\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathrm{XA}\gamma + \boldsymbol{\varepsilon} = \left(\mathrm{XA}\right)\boldsymbol{\gamma} + \boldsymbol{\varepsilon},
$$

conforme representação na Figura 9. Supondo as matrizes A e X injetivas então

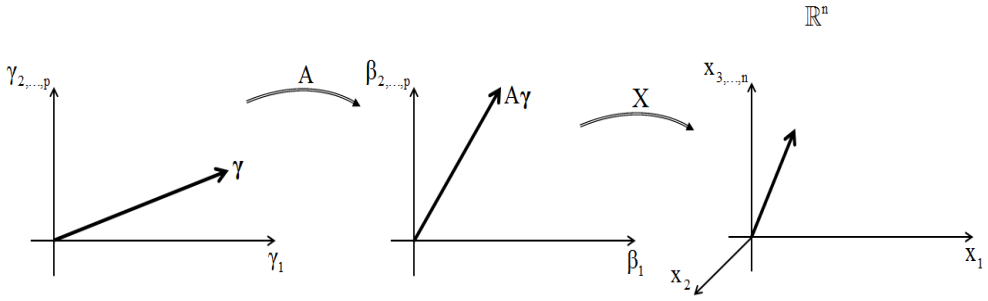


Figura 9 Reparametrização em regressão linear.

$\mathrm{Im}\,(\mathrm{XA}) = \mathrm{Im}\,(\mathrm{X})$. Em relação aos parâmetros dados pelas coordenadas do vetor $\boldsymbol{\gamma}$, a nova matriz da regressão é XA e o estimador de quadrados mínimos para $\boldsymbol{\gamma}$ é dado por

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= \left[(\mathrm{XA})'\mathrm{XA}\right]^{-1}(\mathrm{XA})'\mathbf{y} \\
&= \left[\mathrm{A}'\mathrm{X}'\mathrm{XA}\right]^{-1}\mathrm{A}'\mathrm{X}'\mathbf{y} \\
&= \mathrm{A}^{-1}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\left(\mathrm{A}'\right)^{-1}\mathrm{A}'\mathrm{X}'\mathbf{y} \\
&= \mathrm{A}^{-1}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathbf{y} \\
&= \mathrm{A}^{-1}\hat{\boldsymbol{\beta}}.
\end{aligned}$$

Denominando as novas covariáveis pelo vetor $\mathbf{z}$ e $\mathrm{XA} = \mathrm{Z}_{\mathrm{n}\times\mathrm{p}}$, então

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix}.$$

Observe que os elementos da j-ésima coluna de Z são dados por

$$z_{ij} = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix} \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix} = x_{i1}a_{1j} + x_{i2}a_{2j} + \cdots + x_{ip}a_{pj}$$

e toda uma coluna de Z é então

$$\begin{bmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{nj} \end{bmatrix} = \begin{bmatrix} x_{11}a_{1j} + x_{12}a_{2j} + ... + x_{1p}a_{pj} \\ x_{21}a_{1j} + x_{22}a_{2j} + ... + x_{2p}a_{pj} \\ \vdots \\ x_{n1}a_{1j} + x_{n2}a_{2j} + ... + x_{np}a_{pj} \end{bmatrix}$$

$$= \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} a_{1j} + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} a_{2j} + \cdots + \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix} a_{pj}.$$

Em outras palavras, a j-ésima coluna de Z é uma combinação linear das colunas de X. As constantes da combinação linear são os elementos da j-ésima coluna de A. Logo, uma linha de Z é da forma

$$\mathbf{z}'_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ip} \end{bmatrix} = \begin{bmatrix} a_{11}x_{i1} + a_{21}x_{i2} + \cdots + a_{p1}x_{ip} \\ a_{12}x_{i1} + a_{22}x_{i2} + \cdots + a_{p2}x_{ip} \\ \vdots \\ a_{1p}x_{i1} + a_{2p}x_{i2} + \cdots + a_{pp}x_{ip} \end{bmatrix} = A'\mathbf{x}_i,$$

isto é, no novo modelo linear $\mathbf{y} = (XA)\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, as colunas de Z são novas covariáveis, dadas por combinações lineares das covariáveis X originais. Segue então um resultado bastante importante: o valor predito não se altera quando se faz

essa mudança linear nas covariáveis, ou seja, nas covariáveis originais, a equação de predição é $\hat{\mathrm{y}} = \hat{\boldsymbol{\beta}}'\mathrm{x}$, enquanto que nas novas covariáveis a equação de predição é dada por

$$\begin{aligned}
\hat{\mathrm{y}} &= \hat{\boldsymbol{\gamma}}'\mathbf{z} \\
&= \left(\mathrm{A}^{-1}\hat{\boldsymbol{\beta}}\right)'\mathrm{A}'\mathbf{x} \\
&= \hat{\boldsymbol{\beta}}'\left(\mathrm{A}^{-1}\right)'\mathrm{A}'\mathbf{x} \\
&= \hat{\boldsymbol{\beta}}'\left(\mathrm{A}'\right)^{-1}\mathrm{A}'\mathbf{x} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{x}.
\end{aligned}$$

2.3 Reparametrização na presença de covariáveis categóricas

Quando se tem covariáveis qualitativas, é necessário atribuir valores arbitrários às categorias definidas. Por exemplo, para a covariável estado civil, pode-se atribuir 0 para solteiro e 1 para casado, ou -1 para solteiro e 1 para casado. Como essa arbitrariedade afetará a equação de predição? Suponha por simplicidade que a última coluna da matriz do delineamento seja definida por uma covariável binária assumindo os valores a e b. Fazendo uma reparametrização apenas desta covariável, para uma outra covariável que assuma os valores c e d, respectivamente aos valores a e b, tem-se que se obter uma matriz de forma que

$$\begin{pmatrix}
1 & \mathrm{x}_{1\,2} & \cdots & \mathrm{x}_{1\,\mathrm{p}-2} & a \\
1 & \mathrm{x}_{2\,2} & \cdots & \mathrm{x}_{2\,\mathrm{p}-2} & b \\
\vdots & \vdots & \ddots & & \vdots \\
1 & \mathrm{x}_{\mathrm{n}\,2} & \cdots & \mathrm{x}_{\mathrm{n}\,\mathrm{p}-2} & a
\end{pmatrix}_{\mathrm{n}\times\mathrm{p}}
\begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & x \\
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 \\
0 & 0 & 0 & \cdots & 0 & y
\end{pmatrix}_{p\times p} =$$

$$= \begin{pmatrix} 1 & \mathrm{x}_{1\,2} & \cdots & \mathrm{x}_{1\,\mathrm{p}-2} & c \\ 1 & \mathrm{x}_{2\,2} & \cdots & \mathrm{x}_{2\,\mathrm{p}-2} & d \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}\,2} & \cdots & \mathrm{x}_{\mathrm{n}\,\mathrm{p}-2} & c \end{pmatrix}_{\mathrm{n}\times\mathrm{p}} = \begin{pmatrix} 1 & \mathrm{x}_{1\,2} & \cdots & \mathrm{x}_{1\,\mathrm{p}-2} & \mathrm{x}+a\mathrm{y} \\ 1 & \mathrm{x}_{2\,2} & \cdots & \mathrm{x}_{2\,\mathrm{p}-2} & \mathrm{x}+b\mathrm{y} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}\,2} & \cdots & \mathrm{x}_{\mathrm{n}\,\mathrm{p}-2} & \mathrm{x}+a\mathrm{y} \end{pmatrix}_{\mathrm{n}\times\mathrm{p}}.$$

Portanto, se tem o sistema $\begin{cases} \mathrm{x}+\mathrm{ya}=\mathrm{c} \\ \mathrm{x}+\mathrm{yb}=\mathrm{d} \end{cases}$, cuja solução é $\mathrm{y}=\frac{\mathrm{c}-\mathrm{d}}{\mathrm{a}-\mathrm{b}}$ e $\mathrm{x}=\frac{\mathrm{ad}-\mathrm{cb}}{\mathrm{a}-\mathrm{b}}$.

O novo vetor de estimativa dos parâmetros, relativo às novas covariáveis, é então $\hat{\boldsymbol{\gamma}} = \mathrm{K}^{-1}\hat{\boldsymbol{\beta}}$. De forma a se determinar a inversa K^{-1} será utilizado a fórmula Sherman-Morrison-Woodbury (RENCHER, 2008, p.23): Dada a matriz particionada $\mathrm{A} = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{bmatrix}$, sua inversa é dada por

$$\mathrm{A}^{-1} = \begin{bmatrix} \big(\mathrm{A}_{11}-\mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{A}_{21}\big)^{-1} & -\mathrm{A}_{11}^{-1}\mathrm{A}_{12}\big(\mathrm{A}_{22}-\mathrm{A}_{21}\mathrm{A}_{11}^{-1}\mathrm{A}_{12}\big)^{-1} \\ -\mathrm{A}_{22}^{-1}\mathrm{A}_{21}\big(\mathrm{A}_{11}-\mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{A}_{21}\big)^{-1} & \big(\mathrm{A}_{22}-\mathrm{A}_{21}\mathrm{A}_{11}^{-1}\mathrm{A}_{12}\big)^{-1} \end{bmatrix}.$$

Para a matriz K, $\mathrm{A}_{11}=\mathrm{I}$, A_{21} é o vetor nulo, $\mathrm{A}'_{12} = \begin{bmatrix} \mathrm{x} & 0 & \cdots & 0 \end{bmatrix}$ e $\mathrm{A}_{22}=\mathrm{y}$, e portanto

$$\mathrm{K}^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \frac{\mathrm{x}}{\mathrm{y}} \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{\mathrm{y}} \end{pmatrix}_{p\times p} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \frac{\mathrm{cb}-\mathrm{ad}}{\mathrm{c}-\mathrm{d}} \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \frac{\mathrm{a}-\mathrm{b}}{\mathrm{c}-\mathrm{d}} \end{pmatrix}.$$

Segue então que o vetor estimado $\hat{\boldsymbol{\gamma}}$ para o sistema com a nova covariável difere de $\hat{\boldsymbol{\beta}}$ apenas no intercepto e na componente relativa à covariável original. Neste trabalho a aplicação a dados reais utiliza uma covariável categórica codificada da

forma homozigoto recessivo (0), heterozigoto (1) e homozigoto dominante (2). Neste caso, pode-se recodificar a covariável por a e b quaisquer, porém o terceiro valor deve obedecer a relação linear expressa no sistema anterior.

3 Teoria geral da estimação

O processo de estimação será abordado em um contexto bastante geral e com forte ênfase na geometria, conforme em Kato (2009) . Seja $\mathbf{y}$ um vetor aleatório n-dimensional com distribuição n-variada qualquer. Considere também $\boldsymbol{\mu}$ um vetor n-dimensional de parâmetros desta distribuição. No caso mais relevante, este último é o vetor de médias $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$. A estimação de $\boldsymbol{\mu}$ será obtida pelo processo descrito a seguir. Seja K um subconjunto de $\mathbb{R}^{\mathrm{n}}$ onde, por hipótese, será suposto estar o vetor $\boldsymbol{\mu}$. Portanto, nesse sentido o subconjunto K é o modelo estatístico. Uma vez observado o vetor $\mathbf{y}$, a estimativa de $\boldsymbol{\mu}$ será obtida simplesmente projetando $\mathbf{y}$ no subconjunto K. Tal projeção será denominada $\hat{\boldsymbol{\mu}} = \mu(\mathbf{y})$, e portanto $\mu : \mathbb{R}^{\mathrm{n}} \mapsto \mathrm{K} \subseteq \mathbb{R}^{\mathrm{n}}$. Este esquema, em sua generalidade, certamente apresentará problemas na obtenção das propriedades da função μ. Serão supostas então algumas hipóteses matemáticas. Será exigido que o subconjunto K seja fechado e que a projeção seja obtida pela minimização da distância

$$\mu(\mathbf{y}) = \underset{\mathrm{k}\in\mathrm{K}}{\arg\min} \left\{ \mathrm{d}(\mathbf{y}, \mathrm{k}), \mathrm{k} \in \mathrm{K} \right\},$$

em que a distância d, em geral, é a distância euclideana. Um dos problemas que a função μ pode apresentar é não estar bem definida no sentido de que pode haver mais de um vetor $\mathrm{k} \in \mathrm{K}$ que minimiza a distância, ou simplesmente não existir um ponto que minimize a distância. Faz-se necessário então considerar uma maior restrição do subconjunto K para que a função μ seja bem definida.

Para que se possa garantir a existência de elementos no subconjunto K que minimizem a distância, supõe-se K um subconjunto fechado e convexo. Lembrando que K é convexo se, dados $\mathrm{k}_1, \mathrm{k}_2 \in \mathrm{K}$, então o vetor $\mathrm{t}\,\mathrm{k}_1 + (1 - \mathrm{t})\mathrm{k}_2$, com $0 \leq \mathrm{t} \leq 1$, também pertence à K. Portanto, a menos que se diga o contrário, será

suposto K como sendo fechado e convexo.

Proposição 4. *Se* K *é convexo e fechado, então a projeção* $\mu : \mathrm{R^n} \mapsto \mathrm{K}$*, em que* $\mu(\mathbf{y}) = \arg\min_{\mathrm{K}} \{\mathrm{d}(\mathbf{y}, \mathbf{k}), \mathrm{k} \in \mathrm{K}\}$*, está bem definida.*

Demonstração. Sejam k_1 e k_2 pontos de K que minimizam a distância ao vetor $\mathbf{y}$. Logo, $d(\mathbf{y}, \mathrm{k}_1) = d(\mathbf{y}, \mathrm{k}_2)$. Observando que o triângulo ABC da Figura 10 é

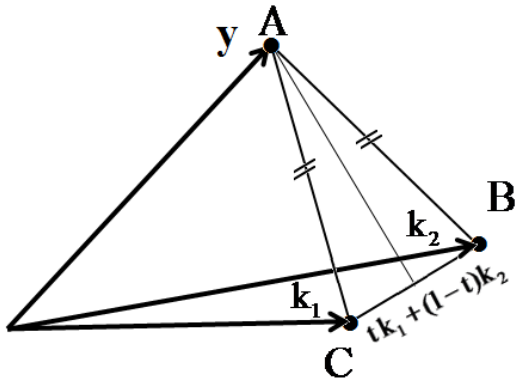


Figura 10 Demonstração da unicidade da projeção em convexos.

isósceles e portanto $\mathrm{d}(\mathrm{t}\,\mathrm{k}_1 + (1-\mathrm{t})\mathrm{k}_2, \mathbf{y}) < \mathrm{d}(\mathrm{k}_2, \mathbf{y})$, tem-se um absurdo, pois pela convexidade de K, $\mathrm{t}\mathbf{k}_1 + (1-\mathrm{t})\mathbf{k}_2 \in \mathrm{K}$.

A existência do vetor em K que minimiza a distância segue do fato de K ser fechado e não será aqui demonstrada. □

Uma observação interessante é que em algumas situações o problema de se minimizar a distância do vetor de dados $\mathbf{y}$ ao modelo K pode ser resolvido como um problema de minimização no espaço paramétrico. Suponha o caso de regressão $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ com $\mathrm{X} : \mathbb{R}^{\mathrm{p}} \mapsto \mathbb{R}^{\mathrm{n}}$ injetiva e que $\mathrm{K} \subseteq \mathrm{Im(X)}$ (Figura 11). Para $\mathrm{k} \in \mathrm{K}$ e $\mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y})$ considere o triângulo da Figura 12. Minimizar a norma da hipotenusa $\mathbf{y} - \mathbf{k}$ é equivalente à minimizar a norma do cateto $\left(\mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y})\right) - \mathrm{k}$, pois a norma do outro cateto não depende do convexo K mas apenas da distância da projeção

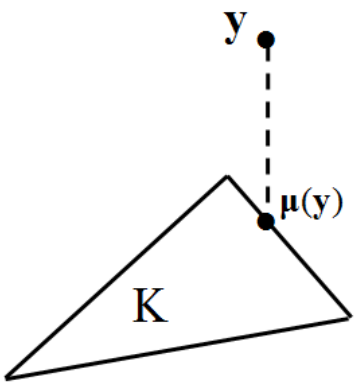


Figura 11 Projeção do vetor $\mathbf{y}$ no convexo K.

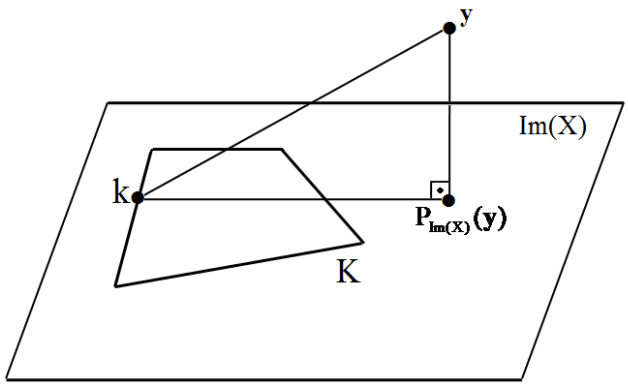


Figura 12 Relação entre $\mathrm{d}\left(\mathbf{y}, \mathbf{k}\right)$ e $\mathrm{d}\left(\mathbf{y}, \mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y})\right)$.

ortogonal e não do ponto k. Como por hipótese $\mathrm{X}: \mathbb{R}^{\mathrm{p}} \mapsto \mathrm{Im}(\mathrm{X}) \subset \mathbb{R}^{\mathrm{n}}$ é injetiva e $\mathrm{K} \subset \mathrm{Im}(\mathrm{X})$, tomando a pré-imagem $\mathrm{K_p} = \mathrm{X}^{-1}(\mathrm{K})$ e observando que se o produto interno de $\mathbb{R}^{\mathrm{p}}$ é definido por $< \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 >_{\mathrm{p}} = \boldsymbol{\beta'}_1 \mathrm{X'X} \boldsymbol{\beta}_2$, então a transformação X é uma isometria sobre a imagem, isto é, $< \mathrm{X}\boldsymbol{\beta}_1, \mathrm{X}\boldsymbol{\beta}_2 >= \boldsymbol{\beta'}_1 \mathrm{X'X} \boldsymbol{\beta}_2 = <\boldsymbol{\beta}_1, \boldsymbol{\beta}_2>_{\mathrm{p}}$. Desta forma, X preserva distâncias. Se $\mathbf{x} \in \mathrm{Im}(\mathrm{X})$ e $\mathrm{X}\mathbf{z} = \mathbf{x}$, então a distância de $\mathbf{z}$ à $\mathrm{K_p}$ é igual à distância de $\mathbf{x}$ à K. Segue então que o problema de se encontrar $\mu(\mathbf{y}) \in \mathrm{K}$ mais próximo de $\mathbf{y}$ é equivalente a se encontrar $\hat{\boldsymbol{\beta}}$ em $\mathrm{K_p}$ o mais perto

possível de $\hat{\boldsymbol{\beta}}_{\text{ols}}$ e $\mathrm{X}\hat{\boldsymbol{\beta}} = \mu(\mathbf{y})$ (Figura 13).

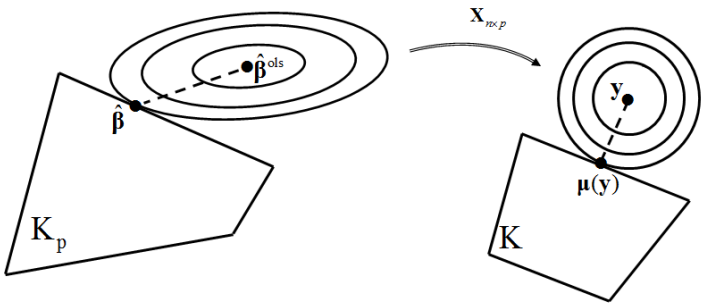


Figura 13 Isometria entre o espaço de dados e o espaço de parâmetros.

Desta forma, $\hat{\boldsymbol{\beta}}$ é a estimativa do vetor de parâmetros obtida pelo processo. As propriedades deste estimador estão claramente vinculadas às propriedades geométricas do conjunto convexo $\mathrm{K_p}$.

3.1 Estimador de quadrados mínimos

Considere o modelo de regressão linear $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ em que X é uma matriz definida pelos valores das covariaveis e $\boldsymbol{\beta}$ um vetor p-dimensional de parâmetros. A matriz X pode ser vista como uma transformação linear do espaço de parâmetros $\mathbb{R}^{\mathrm{p}}$ no espaço de dados $\mathbb{R}^{\mathrm{n}}$, $\mathrm{X} : \mathbb{R}^{\mathrm{p}} \mapsto \mathbb{R}^{\mathrm{n}}$. Como $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}] = \mathrm{E}[\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathrm{X}\boldsymbol{\beta}$, o vetor de médias está contido no subespaço imagem de X, $\mathrm{Im}(\mathrm{X})$, e portanto o modelo K é justamente este subespaço. Observe que este é o caso mais simples possível uma vez que K é fechado e convexo. Mais simples ainda, se tem que a projeção $\mu : \mathbb{R}^{\mathrm{n}} \mapsto \mathrm{Im}(\mathrm{X}) \subseteq \mathbb{R}^{\mathrm{n}}$ é a projeção linear ortogonal. Desta forma, a expressão de $\mu(\mathbf{y})$ pode ser obtida de forma explícita, para o caso de X ser injetiva, como a fórmula matricial $\mu(\mathbf{y}) = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ e $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ (ols - *ordinary least square*). Os estimadores $\mu(\mathbf{y})$ e $\hat{\boldsymbol{\beta}}_{\text{ols}}$ possuem propriedades

ótimas. Este último é não viesado uma vez que

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right] = \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}] = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}[\mathbf{y}] = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

A matriz de covariâncias de $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, pela propriedade $\mathrm{cov}(\mathrm{A}\mathbf{y}) = \mathrm{A}\mathrm{cov}(\mathbf{y})\mathrm{A}'$, é dada por

$$\begin{aligned}
\mathrm{cov}(\hat{\boldsymbol{\beta}}_{\mathrm{ols}}) &= \mathrm{cov}((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}) \\
&= \left[\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right]\mathrm{cov}(\mathbf{y})\left[\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right]' \\
&= \left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathbf{y}\mathbf{y}'\left[\mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\right] \\
&= \left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\sigma^2\mathrm{I}\mathrm{X}\left(\mathrm{X}'\mathrm{X}\right)^{-1} \\
&= \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{I}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} \\
&= \sigma^2(\mathrm{X}'\mathrm{X})^{-1}.
\end{aligned}$$

Nesse contexto, a variância total é

$$\mathrm{var}_{tot}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right) = \sigma^2 tr((\mathrm{X}'\mathrm{X})^{-1}) = \sigma^2\sum_{i=1}^{\mathrm{p}}\frac{1}{\lambda_i},$$

em que λ_i são os autovalores de $(\mathrm{X}'\mathrm{X})$. Pelo teorema de Gauss-Markov, $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ possui variância total mínima dentre todos os estimadores lineares não viesados. A geometria de estimação de quadrados mínimos está descrita na Figura 14.

O estimador de quadrados mínimos obtido acima, a partir de uma consideração puramente geométrica, é obtido a partir da menor distância de um vetor $\mathbf{y}$ a um subespaço, que é a projeção ortogonal. É possível também se deduzir analiticamente a expressão do estimador, minimizando a distância de $\mathbf{y}$ à $\mathrm{Im}(\mathrm{X})$ por

$$\boldsymbol{\mu}(\mathbf{y}) = \arg\min \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$$

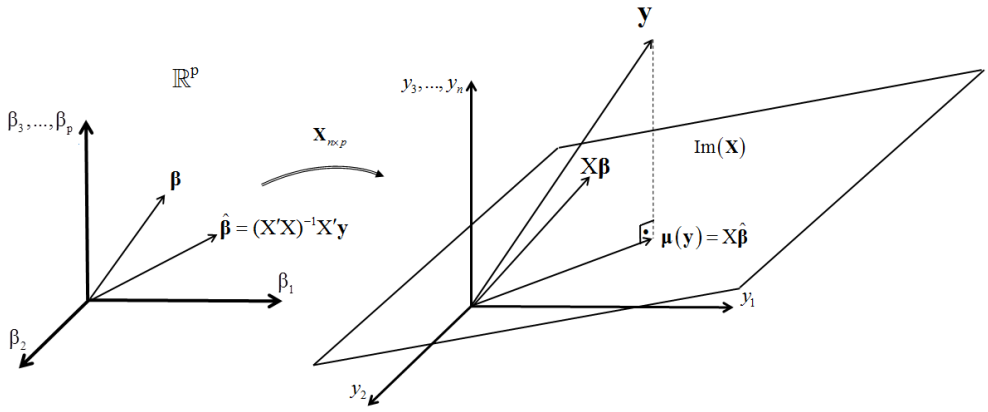


Figura 14 O Estimador de quadrados mínimos.

$$
\begin{aligned}
&= \arg\min < \mathbf{y} - \mathrm{X}\boldsymbol{\beta}, \mathbf{y} - \mathrm{X}\boldsymbol{\beta} > \\
&= \arg\min \{< \mathbf{y}, \mathbf{y} > -2 < \mathbf{y}, \mathrm{X}\boldsymbol{\beta} > + < \mathrm{X}\boldsymbol{\beta}, \mathrm{X}\boldsymbol{\beta} >\}
\end{aligned}
$$

que em termos matriciais pode ser expresso por

$$\boldsymbol{\mu}(\mathbf{y}) = \arg\min \left\{\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta}\right\}.$$

Derivando esta equação e igualando a zero obtém-se as equações normais $\mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}'\mathbf{y}$. No caso em que a matriz X é de posto completo, $\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$. É adequado, como tem-se o termo $\boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta}$, considerar no espaço paramétrico um produto interno modificado, definido por

$$< \boldsymbol{\beta}, \boldsymbol{\beta} >_{\mathrm{p}} = \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta}.$$

Este produto interno introduz em $\mathbb{R}^{\mathrm{p}}$ uma geometria (métrica) modificada. Por exemplo, os vetores $\boldsymbol{\beta}$ com $\|\boldsymbol{\beta}\|_{\mathrm{p}} = \mathrm{c}$, isto é, a esfera de raio c centrada na origem, é de fato um elipsoide pois

$$\|\boldsymbol{\beta}\| = \mathrm{c} \Rightarrow \|\boldsymbol{\beta}\|^2 = \mathrm{c}^2 \Rightarrow \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta} = c^2.$$

Para uma melhor descrição desse elipsoide, como $\mathrm{X}'\mathrm{X}$ é simétrica e suposta de posto completo, seja $\{\mathbf{v}_1, \ldots, \mathbf{v}_\mathrm{p}\}$ uma base ortonormal formada de autovetores de $\mathrm{X}'\mathrm{X}$, ou seja, $\mathrm{X}'\mathrm{X}\mathbf{v}_\mathrm{i} = \lambda_\mathrm{i}\mathbf{v}_\mathrm{i}$. Defina a transformação ortogonal $\mathrm{A}(\mathbf{e}_\mathrm{i}) = \mathbf{v}_\mathrm{i}$, em que $\{\mathbf{e}_1, \ldots, \mathbf{e}_\mathrm{p}\}$ é a base canônica de $\mathbb{R}^\mathrm{p}$. Fica então definida a reparametrização $\tilde{\boldsymbol{\beta}} = \mathrm{A}\boldsymbol{\beta}$, o qual implica que

$$\begin{aligned} \mathrm{c} &= <\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}>_\mathrm{p} \\ &= \tilde{\boldsymbol{\beta}}'\mathrm{X}'\mathrm{X}\tilde{\boldsymbol{\beta}} \\ &= (\mathrm{A}\boldsymbol{\beta})'(\mathrm{X}'\mathrm{X})\mathrm{A}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}'\mathrm{A}'(\mathrm{X}'\mathrm{X})\mathrm{A}\boldsymbol{\beta}. \end{aligned}$$

Como

$$\begin{aligned} \mathrm{A}'(\mathrm{X}'\mathrm{X})\mathrm{A}(\mathbf{e}_\mathrm{i}) &= \mathrm{A}'(\mathrm{X}'\mathrm{X})(\mathbf{v}_\mathrm{i}) \\ &= \mathrm{A}'(\lambda_\mathrm{i}\mathbf{v}_\mathrm{i}) \\ &= \lambda_\mathrm{i}\mathrm{A}'(\mathbf{v}_\mathrm{i}) \\ &= \lambda_\mathrm{i}\mathrm{A}^{-1}(\mathbf{v}_\mathrm{i}) \\ &= \lambda_\mathrm{i}\mathbf{e}_\mathrm{i} \end{aligned}$$

e a matriz $\mathrm{A}'(\mathrm{X}'\mathrm{X})\mathrm{A}$ é da forma diagonal $(\lambda_1, \ldots, \lambda_\mathrm{k})$, então com estes novos parâmetros se tem

$$\begin{aligned} < \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}>_\mathrm{p} &= <\mathrm{A}\boldsymbol{\beta}, \mathrm{A}\boldsymbol{\beta}>_\mathrm{p} \\ &= \boldsymbol{\beta}'(\mathrm{A})'(\mathrm{X}'\mathrm{X})\mathrm{A}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}'\mathrm{A}'(\mathrm{X}'\mathrm{X})\mathrm{A}\boldsymbol{\beta} \\ &= \sum_{i=1}^{\mathrm{k}} \lambda_\mathrm{i}\boldsymbol{\beta}_\mathrm{i}^2 = c. \end{aligned}$$

Esta equação define um elipsoide com eixos nos eixos coordenados e tamanhos λ_i. Uma vez que A é uma transformação ortogonal e portanto uma isometria em relação à métrica usual, esta preserva a forma (Figura 15) e segue que
$c = <\boldsymbol{\beta}, \boldsymbol{\beta}>_p$ é um elipsoide com eixos definidos pelos vetores $\mathbf{v}_i$ e tamanho λ_i.

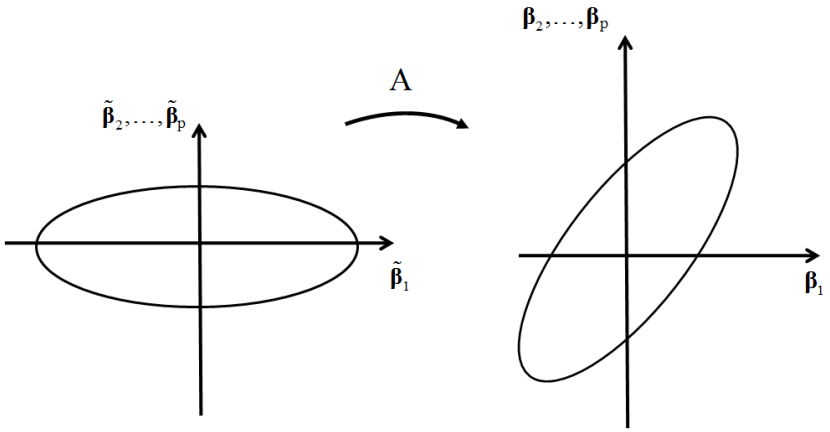


Figura 15 Transformações ortogonais preservam elipsoides.

Uma questão natural é como é a imagem deste elipsoide no subespaço $\mathrm{Im}(X)$. Se $c = <\boldsymbol{\beta}, \boldsymbol{\beta}>_p$, então sua norma em $\mathbb{R}^n$ é

$$\|X\boldsymbol{\beta}\|^2 = < X\boldsymbol{\beta}, X\boldsymbol{\beta} > = \boldsymbol{\beta}' X' X \boldsymbol{\beta} = c,$$

em que $< \ , \ >$ é o produto usual no espaço $\mathbb{R}^n$. Portanto, a imagem do elipsoide é uma esfera p-dimensional no subespaço $\mathrm{Im}(X)$ (Figura 16). Prova-se, através do Teorema de Decomposição Espectral para matrizes, que também vale o resultado inverso. A imagem de uma esfera em $\mathbb{R}^p$ é um elipsoide em $\mathrm{Im}(X)$ (Figura 17). Esta construção geométrica será fundamental para a abordagem que será feita para os demais estimadores tratados mais adiante neste trabalho.

Apesar do estimador de quadrados mínimos ser o estimador linear não viesado de variância mínima (BLUE), a estimação de quadrados mínimos apre-

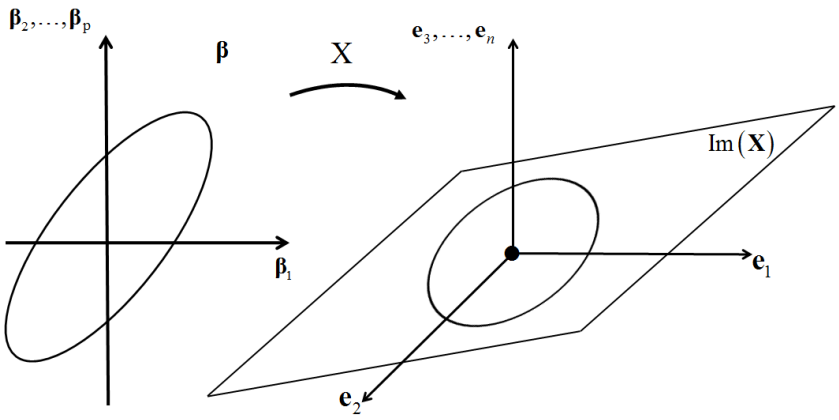


Figura 16 Transformações de elipsoides em esferas.

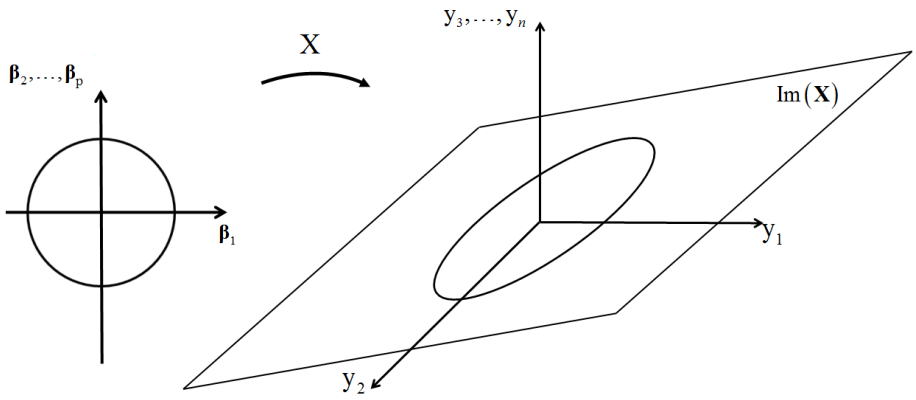


Figura 17 Transformação de esferas em elipsoides.

senta problemas em pelo menos duas situações. Uma delas é quando ocorre quasi-colinearidade, que pode ocorrer quando se tem, por exemplo, pelo menos duas covariáveis altamente correlacionados. Nesse caso, com alta probabilidade, pode-se ter que as duas colunas relativas a estas covariáveis na matriz de delineamento X sejam muito próximas. Tal fato acarreta que pelo menos um dos autovalores da matriz $\mathrm{X}'\mathrm{X}$ é próximo de zero. Como a variância total do estimador de quadrados mínimos $\hat{\boldsymbol{\beta}}_{\text{ols}}$, dada pelo traço da matriz de covariâncias, $\text{cov}\left(\hat{\boldsymbol{\beta}}_{\text{ols}}\right) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$,

depende dos inversos dos autovalores, esta poderá ser excessiva, inviabilizando seu uso. A outra situação ocorre quando $p > n$, isto é, o número de covariáveis é maior do que o número de dados, situação comum em estudos genéticos. Neste caso, a matriz $\mathrm{X}'\mathrm{X}$ é singular, o que leva à necessidade da utilização de inversas generalizadas.

Qual seria uma alternativa ao método de estimação de quadrados mínimos? Ora, tal método é baseado na distância mínima. Outra possibilidade é permitir a utilização de uma distância que não seja a mínima. Esse procedimento recebe o nome genérico de método de quadrados mínimos penalizado. No sentido de explicitar as várias penalizações usualmente utilizadas, será então abordada em mais detalhes a geometria da regressão linear.

Uma vez observado o vetor de dados $\mathbf{y}$, considere no subespaço $\mathrm{Im}\,(\mathrm{X})$ todos os vetores $\mathrm{X}\boldsymbol{\beta}$ que estão a uma distância d do vetor $\mathbf{y}$. Para visualizar os vetores com essa propriedade, recorre-se ao triângulo fundamental.

$$\begin{aligned} d^2 &= \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 \\ &= \left\|\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} + \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \mathrm{X}\boldsymbol{\beta}\right\|^2 \\ &= \left\|\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right\|^2 + \left\|\mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \mathrm{X}\boldsymbol{\beta}\right\|^2, \end{aligned}$$

em que a segunda igualdade se justifica pelo fato do vetor $\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ ser perpendicular ao subespaço $\mathrm{Im}\,(\mathrm{X})$. Como $\left\|\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right\|$ é um valor fixo, tem-se que os vetores $\mathrm{X}\boldsymbol{\beta}$ satisfazem a equação $\left\|\mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \mathrm{X}\boldsymbol{\beta}\right\| = \mathrm{r}$, com r constante, isto é, forma uma hiperesfera de raio r no subespaço $\mathrm{Im}\,(\mathrm{X})$, como descrito na Figura 18.

Os vetores nesta hiperesfera, do ponto de vista de distância aos dados, são indistintos. Portanto, o processo de estimação, isto é, escolha de um vetor de parâmetros $\boldsymbol{\beta}$ particular deve seguir algum tipo de restrição. Novamente, a chave

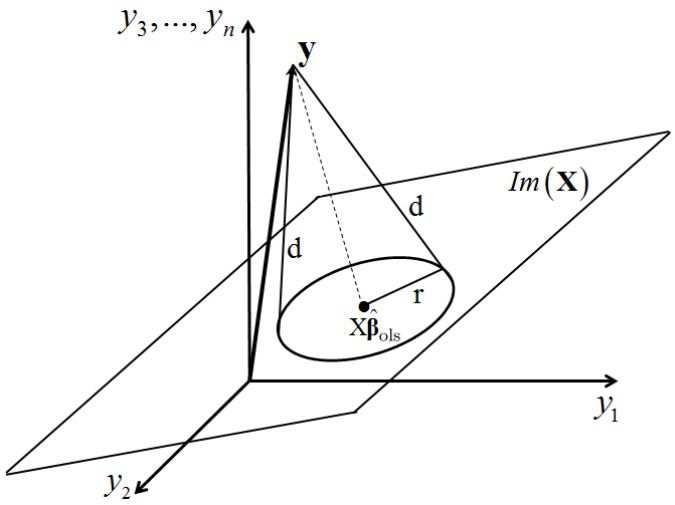


Figura 18 Vetores equidistantes à $\mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$

para se entender os processos de estimação das teorias acima citadas é a geometria. Que propriedades possuem os vetores $\boldsymbol{\beta}$ que são levados pela transformação linear X nessa hiperesfera?

Observando que $\|v\|^2 = v'v$, tem-se

$$\begin{aligned} r^2 &= \left\|\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2 \\ &= \left(\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right) \\ &= \left(\mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right)' \left(\mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \\ &= \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathrm{X}'\mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \end{aligned}$$

Portanto, os vetores $\boldsymbol{\beta}$ que são levados na hiperesfera formam um elipsoide centrado em $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, conforme representado na Figura 19.

Variando-se o raio r, varia-se de forma equivalente o elipsoide sem no entanto alterar as direções de seus eixos principais e sua excentricidade (Figura 20). Nesse sentido, o raio r pode ser considerado como um parâmetro de ajuste

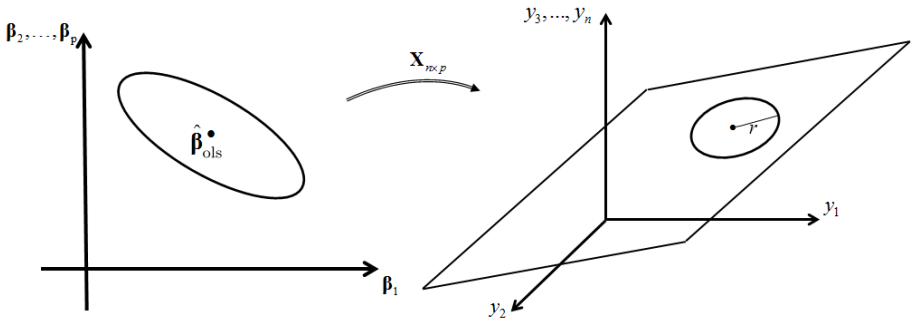


Figura 19 Relação entre a hiperesfera e o elipsoide.

para o processo de estimação.

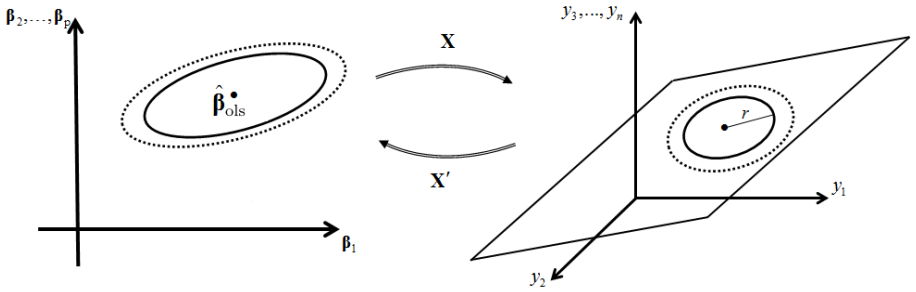


Figura 20 Variação do raio da hiperesfera.

4 A regressão Ridge

Um procedimento usual na estatística é a abordagem conservadora de, entre duas estimativas igualmente plausíveis, se optar pela de menor tamanho ou de menor norma. Utilizando essa filosofia, fixado o valor de r, deve-se escolher na elipse centrada em $\hat{\boldsymbol{\beta}}_{\text{ols}}$ o vetor $\boldsymbol{\beta}$ de menor norma, que será denominado $\hat{\boldsymbol{\beta}}_R(r)$ (Figura 21). Esse procedimento de estimação é denominado estimação Ridge (HOERL; KENNARD 1970),

Dessa forma, o estimador Ridge $\hat{\boldsymbol{\beta}}_R(r)$ é um estimador de encolhimento no sentido que $\left\|\hat{\boldsymbol{\beta}}_R(r)\right\| < \left\|\hat{\boldsymbol{\beta}}_{\text{OLS}}\right\|$ e também, evidentemente, um estimador

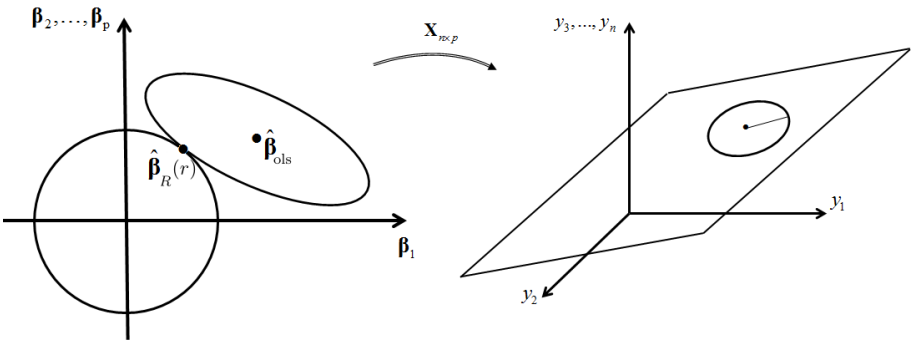


Figura 21 O estimador ridge.

viesado.

Uma das grandes vantagens do método de estimação Ridge é que, variando-se o parâmetro r obtém-se uma curva $\hat{\boldsymbol{\beta}}_R(r)$ no espaço de parâmetros. Tomando-se as coordenadas $\left(\hat{\boldsymbol{\beta}}_R\right)_i(r)$ de $\hat{\boldsymbol{\beta}}_R(r)$, é possível uma descrição bidimensional do comportamento multidimensional do estimador, traçando-se simultaneamente os gráficos de $\left(\hat{\boldsymbol{\beta}}_R\right)_i(r)$ em função do parâmetro r. Tais gráficos são denominados *ridgetrace*, e permitem uma análise gráfica bastante útil do processo de estimação. A Figura 22 apresenta um destes *ridgetraces*.

A questão de se determinar um valor ótimo para o parâmetro r, isto é, o quanto se dever ter de encolhimento para que seja mínimo o erro quadrático médio $\mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\hat{\boldsymbol{\beta}}_{\mathrm{R}}(r)-\boldsymbol{\beta}\right\|^2\right]$, é uma das questões mais estudadas na teoria, envolvendo problemas analíticos complexos. Tal valor depende do vetor de parâmetro populacional $\boldsymbol{\beta}$, que deve ser estimado obtendo-se estimadores do valor ótimo para r bastante complexos. Uma forma gráfica de se estimar uma região para o valor ótimo de r é o intervalo em que as curvas de *ridgetrace* se tornam aproximadamente paralelas ao eixo das abscissas.

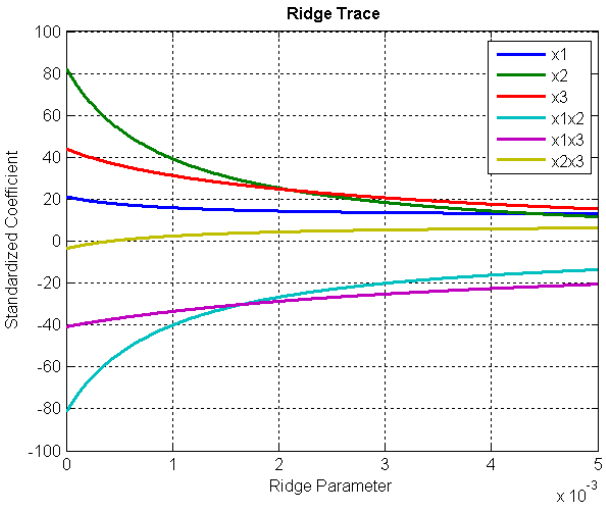


Figura 22 Exemplo de *ridgetrace* em $\mathbb{R}^6$.

4.1 O estimador Ridge

Nesta subseção será explicitada a obtenção do estimador $\hat{\boldsymbol{\beta}}_R(r)$.

O estimador de mínimos quadrados, pelo teorema de Gauss-Markov, é o melhor estimador linear não viesado (BLUE). Uma pergunta natural que se pode fazer é, existe algum estimador viesado com erro quadrático menor? Neste sentido, a regressão de encolhimento trata de estimadores que, ao se permitir viés, fornecem estimativas melhores no sentido de se diminuir o erro quadrático médio. Para tratarem o problema de quasi-colinearidade, Hoerl; Kennard (1970) definiram um estimador denominado Estimador Ridge utilizando uma penalização no método de quadrados mínimos. Considere o problema de regressão linear $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Novamente, o modelo K é o subespaço $\mathrm{Im}(\mathrm{X})$. O estimador de quadrados mínimos é obtido ao se projetar ortogonalmente o vetor $\mathbf{y}$ em $\mathrm{Im}(\mathrm{X})$. Introduz-se então uma penalização afirmando que as estimativas possíveis estão a uma distância r da projeção ortogonal, isto é, estão em uma esfera de raio r contida em $\mathrm{Im}(\mathrm{X})$, esfera

esta centrada na estimativa de quadrados mínimos (Figura 23). Esta penalização gera uma problema, uma vez que em razão da projeção ser ortogonal, todos os pontos desta esfera estão à mesma distância de $\mathbf{y}$. Como escolher então uma estimativa? A ideia é a seguinte: os vetores $\mathbf{z}$ que pertencem a esta esfera satisfazem a condição $< \mathbf{z} - \mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y}), \mathbf{z} - \mathrm{P}_{\mathrm{Im(X)}}(\mathbf{y}) >= \mathrm{r}^2$. Considere então os vetores $\boldsymbol{\beta}$ no espaço paramétrico tais que $\mathrm{X}\boldsymbol{\beta} = \mathbf{z}$. Logo,

$$\begin{aligned}
\mathrm{r}^2 &= < \mathbf{z} - \mathrm{P}_{\mathrm{Im(X)}}\mathbf{y}, \mathbf{z} - \mathrm{P}_{\mathrm{Im(X)}}\mathbf{y} > \\
&= < \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ols}, \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}_{ols} > \\
&= < \mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right), \mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right) > \\
&= \left[\mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right)\right]' \mathrm{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right) \\
&= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols})' \mathrm{X}' \mathrm{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}),
\end{aligned}$$

de onde segue que a pré-imagem da esfera é um elipsoide centrado em $\hat{\boldsymbol{\beta}}_{ols}$ no espaço paramétrico $\mathbb{R}^{\mathrm{p}}$.

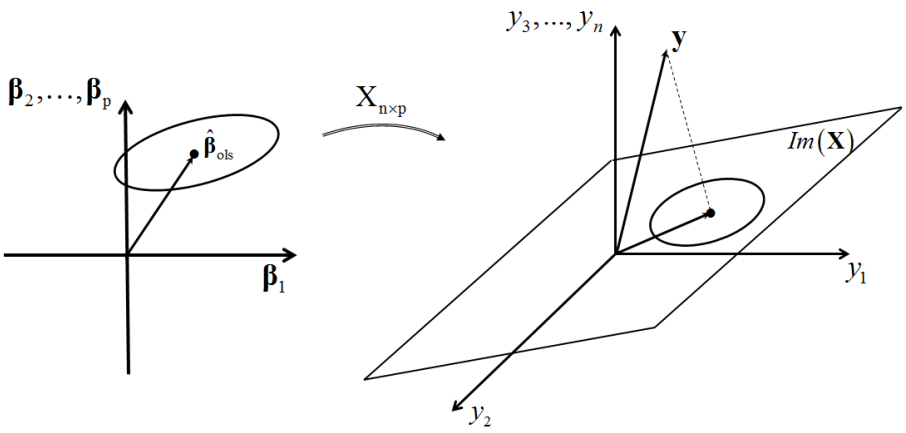


Figura 23 Penalização na definição do estimador Ridge.

Adota-se então uma atitude conservadora. Todos os $\boldsymbol{\beta}$ neste elipsoide são

estimativas viáveis. Opta-se então por aquela de menor norma, isto é, toma-se o $\boldsymbol{\beta}$ obtido como tangente entre o elipsoide e uma esfera centrada na origem. A dedução analítica da expressão deste estimador é como se segue, e é representada pela Figura 24.

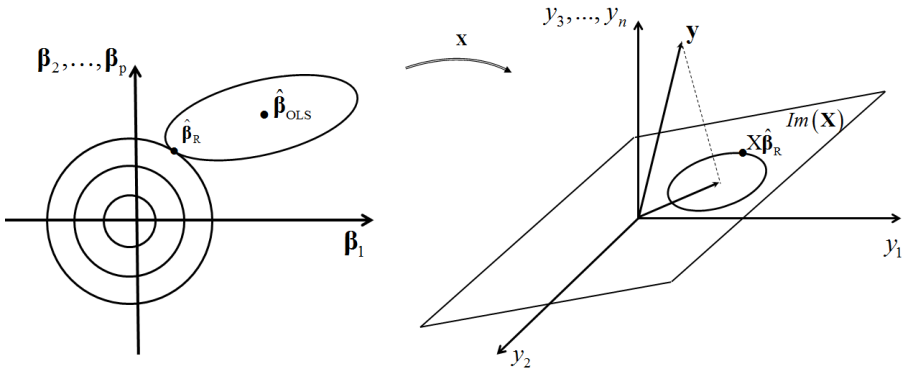


Figura 24 A geometria do estimador Ridge.

Para cada valor fixo de r, a estimativa $\hat{\boldsymbol{\beta}}_R(r)$ é, como descrito geometricamente, obtida como um problema de minimização. Analiticamente este problema pode ser descrito de duas formas equivalentes. Primeiramente no espaço de parâmetros.

Se quer minimizar a função $\min_{\beta} \|\boldsymbol{\beta}\|^2$ sujeito à restrição

$$\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathrm{X}'\mathrm{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) = \mathrm{r}^2.$$

Utilizando o método dos multiplicadores de Lagrange, a Lagrangiana é

$$\begin{aligned} L(\boldsymbol{\beta}, \lambda) &= \|\boldsymbol{\beta}\|^2 + \lambda \left(\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathrm{X}'\mathrm{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) - \mathrm{r}^2 \right) \\ &= \boldsymbol{\beta}'\boldsymbol{\beta} + \lambda \left(\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathrm{X}'\mathrm{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) - \mathrm{r}^2 \right). \end{aligned}$$

Derivando em relação aos parâmetros e igualando à 0 tem-se

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\beta} + \lambda \left[2\mathrm{X}'\mathrm{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) \right] = 0$$

e

$$\frac{\partial L}{\partial \lambda} = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)' \mathrm{X}'\mathrm{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) - r^2 = 0.$$

Logo,

$$\boldsymbol{\beta} + \lambda \mathrm{X}'\mathrm{X}\boldsymbol{\beta} = \lambda \mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}} \Rightarrow$$

$$\left(\mathrm{I} + \lambda \mathrm{X}'\mathrm{X} \right) \boldsymbol{\beta} = \lambda \mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}}.$$

Portanto a solução é dada explicitamente por

$$\hat{\boldsymbol{\beta}}_R(r) = \left(\mathrm{I} + \lambda \mathrm{X}'\mathrm{X} \right)^{-1} \lambda \mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}} = \left(\frac{1}{\lambda}\mathrm{I} + \mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathrm{X}\hat{\boldsymbol{\beta}}.$$

Uma vez que $\hat{\boldsymbol{\beta}} = \left(\mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y}$, então

$$\hat{\boldsymbol{\beta}}_R(r) = \left(\frac{1}{\lambda}\mathrm{I} + \mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y} = \left(\mathrm{kI} + \mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y},$$

sendo $k = \frac{1}{\lambda}$.

O valor de k em função de r é obtido substituindo-se $\hat{\boldsymbol{\beta}}_R(r)$ na restrição

$$\left(\hat{\boldsymbol{\beta}}_R(r) - \hat{\boldsymbol{\beta}} \right)' \mathrm{X}'\mathrm{X} \left(\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\mathrm{r}) - \hat{\boldsymbol{\beta}} \right) = \mathrm{r}^2$$

$$\left(\left(k\mathrm{I} + \mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y} - \left(\mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y} \right)' \mathrm{X}'\mathrm{X} \left(\left(\mathrm{kI} + \mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y} - \left(\mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y} \right) = \mathrm{r}^2$$

Como é fácil atribuir um valor para k e obter o valor correspondente de r, geralmente o estimador Ridge é expresso em função de k no lugar de r.

$$\hat{\boldsymbol{\beta}}_R(\mathrm{k}) = \left(\mathrm{kI} + \mathrm{X}'\mathrm{X} \right)^{-1} \mathrm{X}'\mathbf{y}$$

Tal substituição é adequada, porém k não tem um significado geométrico como o tem r. O problema variacional que define o estimador Ridge admite uma

outra forma equivalente, no espaço de dados. Considere $\boldsymbol{\beta}$ sobre uma esfera de raio r, isto é, $\boldsymbol{\beta'\beta} = r^2$. A imagem desta esfera pela transformação X é uma elipse.

A ideia agora é obter $\boldsymbol{\beta}$ tal que $\mathrm{X}\boldsymbol{\beta}$ esteja o mais próximo possível de $\mathbf{y}$, isto é, $\min_{\beta} \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$ restrito à elipse imagem da esfera $\|\boldsymbol{\beta}\|^2 = r^2$.

A Lagrangeana desse problema é

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

e então,

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\mathbf{y}'\mathrm{X} + 2\mathrm{X}'\mathrm{X}\boldsymbol{\beta} + \lambda(2\boldsymbol{\beta}) = 0 \\
-\mathbf{y}'\mathrm{X} + \left(\mathrm{X}'\mathrm{X} + \lambda\mathrm{I}\right)'\boldsymbol{\beta} = 0 \\
\hat{\boldsymbol{\beta}}_R(r) = \left(\mathrm{X}'\mathrm{X} + \lambda\mathrm{I}\right)^{-1}\mathrm{X}'\mathbf{y}.
\end{aligned}$$

No espaço de dados, tem-se a interpretação

$$\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 = \left\|\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right\|^2 + \left\|\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}}\right\|^2 = cte + \left\|\mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{ols}} - \mathrm{X}\boldsymbol{\beta}\right\|^2.$$

O que se quer então é a minimização

$$\min_{\beta} \left\|\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2 = \min_{\beta} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathrm{X}'\mathrm{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right),$$

sujeito à restrição $\boldsymbol{\beta'\beta} = r^2$, isto é, para os vetores $\boldsymbol{\beta}$ sobre a esfera obter aquele na elipse centrada em $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ de menor tamanho pela métrica de Mahalanobis (Figura 25).

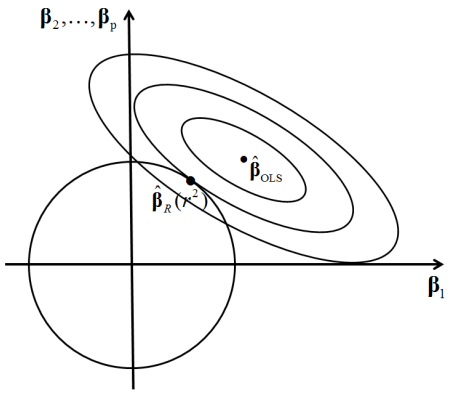


Figura 25 Projeção definida pela métrica de Mahalanobis.

4.2 LASSO

Em regressão linear além do problema de estimação do vetor de parâmetros também se coloca o problema de seleção de covariáveis. Na presença de muitas covariáveis é razoável que o pesquisador queira selecionar apenas algumas delas que mais afetam a variável resposta, isto é, um modelo mais parcimonioso. Uma possibilidade é se obter $\hat{\boldsymbol{\beta}}_{\text{ols}}$ e eliminar as covariáveis relativas às coordenadas de $\hat{\boldsymbol{\beta}}_{\text{ols}}$ de valores relativamente muito menores que as outras. Se o estimador de quadrados mínimos apresentar alta variabilidade (quasi multicolinearidade), este procedimento evidentemente apresenta problemas e é conhecido que não é estável, ou seja, se um novo vetor de respostas $\mathbf{y}$ é observado, com alta probabilidade, covariáveis diferentes serão selecionadas.

Uma alternativa seria a de se obter um processo de estimação que além de apresentar pouca variabilidade, geralmente obtida por algum processo de encolhimento, gere com alta probabilidade estimativas $\hat{\boldsymbol{\beta}}$ em que várias de suas componentes sejam nulas. Desta forma teria-se um processo automático de seleção de covariáveis. Esta é de fato a proposta de um processo de estimação denominado

LASSO, acrônimo de *Least Absolute Shrinkage and Selection Operator*, proposto por Tibshirani (1996). Tal processo é discutido em detalhes neste texto.

Como um dos objetivos deste trabalho é facilitar a leitura do artigo original, será utilizada a mesma notação do artigo. Em particular, $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ será denotado por $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$.

Sejam então os dados $(\mathbf{x}_{\mathrm{i}}, \mathrm{y}_{\mathrm{i}})$, com $\mathrm{i} = 1, \ldots, \mathrm{n}$ para $\mathbf{x}_{\mathrm{i}} = (x_{\mathrm{i}1}, \ldots, x_{\mathrm{i}\,\mathrm{p}})$ sendo $\mathbf{x}_{\mathrm{i}}$ o vetor linha da matriz de delineamento, dado pelos valores das variáveis preditoras (covariáveis) e y_{i} as respostas observadas. Se $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{\mathrm{p}}) \in \mathbb{R}^{\mathrm{p}}$, o estimador LASSO é definido por

$$\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}\right) = \arg\ \min \left\{ \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y}_{\mathrm{i}} - \alpha - \sum_{\mathrm{j}} \beta_{\mathrm{j}} x_{\mathrm{ij}} \right)^2 \right\},$$

sujeito à restrição $\sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_{\mathrm{j}}| \leq \mathrm{t}$.

É fácil ver que $\hat{\alpha} = \bar{\mathbf{y}}$. É adequado por uma transformação de dados colocar o problema na forma normalizada supondo $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{ij}} = 0$, $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{ij}}^2 = 1$ e $\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y}_{\mathrm{i}} = \bar{\mathbf{y}} = 0$. Desta forma, o estimador pode ser escrito como

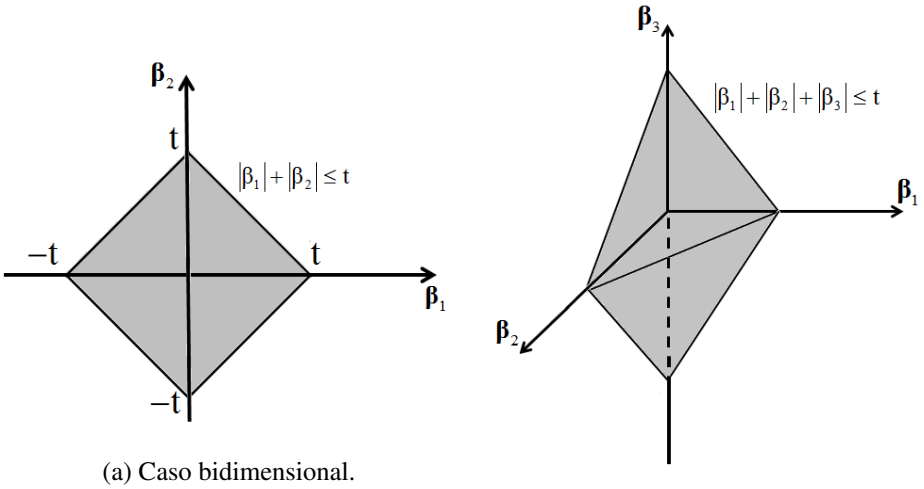
$$\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} = \arg\min \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y}_{\mathrm{i}} - \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{x}_{\mathrm{ij}} \right)^2 \right),$$

ou também pode ser interpretado como $\arg\min \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$ sujeito a restrição $\sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_{\mathrm{j}}| \leq \mathrm{t}$.
Esta última forma fica representada na forma Lagrangeana por

$$\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{p}}}{\arg\min} \left(\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right).$$

Em termos de modelo geométrico, o subconjunto fechado e convexo K_{p},

definido por $\sum_{j=1}^{p} |\beta_j| \leq t$, é um hipercubo cujas diagonais estão sobre os eixos coordenados, e o centro sobre a origem. O valor $t \geq 0$ pode ser considerado como um parâmetro de ajuste. Para o caso bidimensional e tridimensional, K_p é como se segue na Figura26.



(a) Caso bidimensional.

(b) Caso tridimensional.

Figura 26 Restrição K_p LASSO.

No espaço de dados $\mathbb{R}^n$, o convexo $K = X(K_p)$ é semelhante a K_p substituindo os eixos coordenados por eixos definidos pelos vetores das covariáveis $\mathbf{x}_i$. O estimador LASSO $\hat{\boldsymbol{\beta}}^{lasso}$ é então obtido da forma anteriormente descrita. Obtém-se sobre o convexo K o ponto mais próximo do vetor de dados $\mathbf{y}$, o qual será denominado $P_K(\mathbf{y})$. Desta forma, $X\left(\hat{\boldsymbol{\beta}}^{lasso}\right) = P_K(\mathbf{y})$. O ponto $P_K(\mathbf{y})$ também pode ser obtido, conforme descrito anteriormente em termos do triângulo fundamental, como o ponto de K mais próximo de $X\left(\hat{\boldsymbol{\beta}}^{o}\right) = X(X'X)^{-1}X'\mathbf{y}$.

A obtenção de $\hat{\boldsymbol{\beta}}^{lasso}$ também pode ser descrita no espaço paramétrico, uma vez que minimizar a distância do vetor $\mathbf{y}$ ao convexo K é, como também descrito anteriormente, equivalente a se minimizar a distância da estimativa de

mínimos quadrados $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ ao convexo $\mathrm{K_p}$. Entretanto como esta distância é definida pela métrica $< \mathbf{v}_1, \mathbf{v}_2 >_{\mathrm{p}} = \mathbf{v'}_1 \mathrm{X'X} \mathbf{v}_2$, $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{\mathrm{p}}$, a obtenção do ponto de $\mathrm{K_p}$ mais próximo de $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ pode ser descrita como o primeiro ponto de $\mathrm{K_p}$ que tangencia uma elipse centrada em $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$, conforme Figura 27.

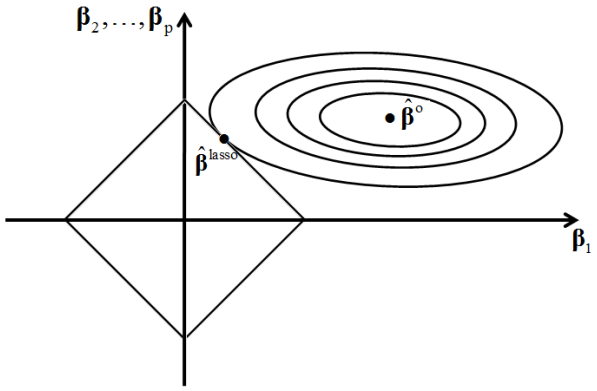


Figura 27 Obtenção do estimador LASSO.

O conjunto convexo $\mathrm{K_p}$ impõe duas restrições que possuem uma justificativa estatística. A primeira delas é manter os valores das estimativas $\left(\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}\right)_{\mathrm{i}}$ limitadas. Esta é uma posição conservadora, para evitar estimativas excessivas. A segunda justificativa é bem mais sofisticada. O convexo $\mathrm{K_p}$ tem como borda uma hipersuperfície formada de planos tais que estas se intersectam em superfícies formadas de planos de dimensões menores, que podemos pensar como "arestas". Como as dimensões das arestas variam, ocorre que os pontos nas arestas possuem algumas coordenadas nulas. É intuitivo pensar que estas arestas têm uma maior probabilidade de conterem o ponto mais próximo de $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$. Portanto, o estimador $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ tem uma maior probabilidade de possuir algumas coordenadas nulas. Neste sentido o estimador funciona também como um processo de seleção de covariáveis e é esta propriedade que mais impulsionou seu estudo e utilização.

Para exemplificar o método, segue um caso particular em que se tem orto-

gonalidade, isto é, $\mathrm{X}'\mathrm{X} = \mathrm{I}$.

Passando então a discutir o caso ortogonal, como exemplo mais simples, serão explicitados os cálculos para a situação de 2 covariáveis. Observe que, como por hipótese $\mathrm{X}'\mathrm{X} = \mathrm{I}$, o produto interno $< \, , >_{\mathrm{p}}$ é o produto usual e portanto no lugar de elipses centradas em $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ tem-se círculos. O estimador LASSO é obtido pela projeção do estimador de quadrados mínimos $\hat{\boldsymbol{\beta}}^{\mathrm{o}} = (\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}, \hat{\boldsymbol{\beta}}_2^{\mathrm{o}})$ no retângulo $|\boldsymbol{\beta}_1| + |\boldsymbol{\beta}_2| \leq \mathrm{t}$. Vamos supor $\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} > 0$, $\hat{\boldsymbol{\beta}}_2^{\mathrm{o}} > 0$ e $\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + \hat{\boldsymbol{\beta}}_2^{\mathrm{o}} > \mathrm{t}$. Uma vez que a projeção ortogonal é a usual, tem-se que os vetores na faixa hachurada, representados na Figura 28, se projetam sobre a aresta, enquanto os vetores fora da faixa hachurada projetam-se no vértice $(0, \mathrm{t})$ ou $(\mathrm{t}, 0)$.

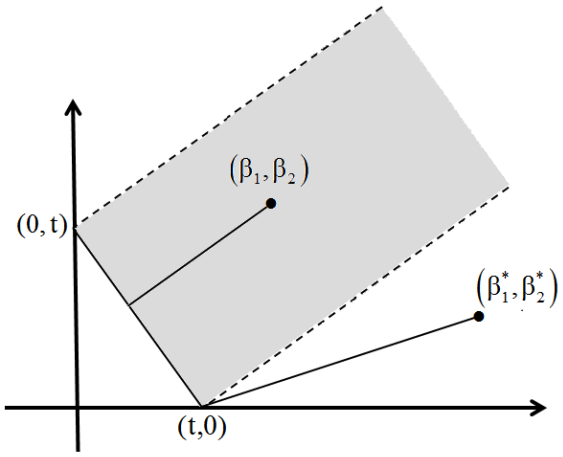


Figura 28 O método de estimação LASSO.

Suponha então que $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ esteja na faixa hachurada. A reta parametrizada que passa por $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ e é ortogonal à aresta do quadrado é $\vec{r}(s) = (\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + s, \hat{\boldsymbol{\beta}}_2^{\mathrm{o}} + s)$, com $s \leq 0$ e $\vec{r}(0) = (\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}, \hat{\boldsymbol{\beta}}_2^{\mathrm{o}})$. Esta reta toca a aresta quando

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + s + \hat{\boldsymbol{\beta}}_2^{\mathrm{o}} + s = \mathrm{t}$$

$$\Rightarrow 2\mathrm{s} = \mathrm{t} - \left(\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}\right)$$
$$\Rightarrow \mathrm{s} = \frac{\mathrm{t}}{2} - \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}\right).$$

A intersecção da reta com a aresta é então

$$\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + \frac{\mathrm{t}}{2} - \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}\right)\ ,\ \hat{\boldsymbol{\beta}}_2^{\mathrm{o}} + \frac{\mathrm{t}}{2} - \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}\right)\right),$$

isto é,

$$\left(\frac{\mathrm{t}}{2} + \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}{2}\ ,\ \frac{\mathrm{t}}{2} - \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}{2}\right).$$

Portanto, o estimador LASSO é dado por

$$\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} = \left(\frac{\mathrm{t}}{2} + \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}{2}\ ,\ \frac{\mathrm{t}}{2} - \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - \hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}{2}\right).$$

Os vetores que se projetam em $(0, \mathrm{t})$ satisfazem a propriedade de que a reta $s = \hat{\boldsymbol{\beta}}^{\mathrm{o}} + s\overrightarrow{\mathbf{1}}$ intersecta a reta definida pela aresta $\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 = \mathrm{t}$ no segundo ou quarto quadrante. Para os vetores $\hat{\boldsymbol{\beta}}$ do primeiro quadrante que não estão na faixa hachurada da Figura 28 ocorre que para algum valor $\tilde{s}$ uma das coordenadas se anula, pois a reta intersecta os eixos coordenados (Figura 29). Esta condição implica que o ponto de $\mathrm{K_p}$ mais próximo de $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ é o ponto $(\mathrm{t}, 0)$ ou $(0, \mathrm{t})$, isto é, $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ se projeta em $(\mathrm{t}, 0)$ ou $(0, \mathrm{t})$.

Para $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ como na Figura 29, tem-se que para algum $\tilde{s}$ a reta $\vec{r}(s) = (\hat{\beta}_1^{\mathrm{o}} - s, \hat{\beta}_2^{\mathrm{o}} - s)$ é tal que

$$\hat{\beta}_2^{\mathrm{o}} - s = 0 \Rightarrow \hat{\beta}_2^{\mathrm{o}} = \tilde{s}.$$

Como neste caso $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ está na região definida pelas retas $(\mathrm{t} + s, s)$ e o eixo positivo, então

$$\hat{\beta}_2^{\mathrm{o}} < \hat{\beta}_1^{\mathrm{o}} - s \Rightarrow s < \hat{\beta}_1^{\mathrm{o}} - \hat{\beta}_2^{\mathrm{o}} \Rightarrow \frac{s}{2} < \frac{\hat{\beta}_1^{\mathrm{o}} - \hat{\beta}_2^{\mathrm{o}}}{2}.$$

Desta forma, o estimador LASSO pode ser dado por uma única relação

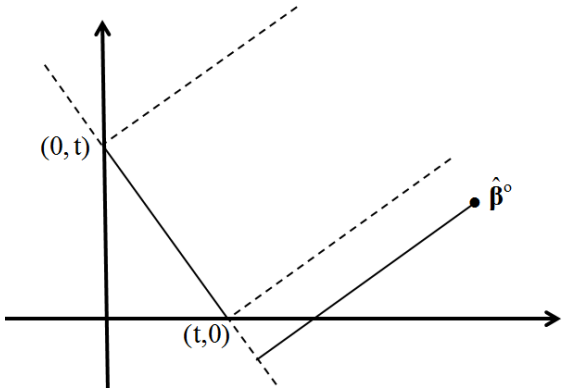


Figura 29 Estimador LASSO no caso bidimensional.

(TIBSHIRANI, 1996, p.272),

$$\hat{\beta}_1^{\mathrm{o}} = \left(\frac{\mathrm{t}}{2} + \frac{\hat{\beta}_1^{\mathrm{o}} - \hat{\beta}_2^{\mathrm{o}}}{2}\right)^+ \text{ e } \hat{\beta}_2^{\mathrm{o}} = \left(\frac{\mathrm{t}}{2} - \frac{\hat{\beta}_1^{\mathrm{o}} - \hat{\beta}_2^{\mathrm{o}}}{2}\right)^+,$$

em que $(x)^+$ significa que $(x)^+ = x$ se $x \geq 0$ e $(x)^+ = 0$ se $x \leq 0$.

Nas situações em que $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ está em outros quadrantes, a análise segue de forma análoga.

Para o caso ortogonal em uma dimensão qualquer, o estimador de quadrados mínimos é da forma

$$\hat{\boldsymbol{\beta}}^{\mathrm{o}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} = \mathrm{X}'\mathbf{y} = \left(\sum_{\mathrm{j}=1}^{\mathrm{n}} x_{\mathrm{j1}}\mathrm{y_j}, \ldots, \sum_{\mathrm{j}=1}^{\mathrm{n}} x_{\mathrm{jk}}\mathrm{y_j}\right).$$

Se $\hat{\boldsymbol{\beta}}^{\mathrm{o}} \in \mathrm{K_p}$ então $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} = \hat{\boldsymbol{\beta}}^{\mathrm{o}}$. Entretanto, se $\hat{\boldsymbol{\beta}}^{\mathrm{o}} \notin \mathrm{K_p}$ a projeção de $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ ocorrerá no bordo de $\mathrm{K_p}$, $(\partial \mathrm{K_p})$, definido por $\sum_{\mathrm{i}=1}^{\mathrm{p}} |\boldsymbol{\beta}_{\mathrm{i}}| = \mathrm{t}$. O vetor normal a $\partial \mathrm{K_p}$ nas faces regulares (de dimensão $\mathrm{p}-1$) é o vetor $\overrightarrow{\mathbf{1}} = (\pm 1, \ldots, \pm 1)$, em que o número de sinais negativos depende da face. Para as faces singulares (intersecção de hiperfaces), o vetor normal é formado por números ± 1 e 0. Para fixar a ideia, suponha que $\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \neq 0, \forall \mathrm{i} = 1, \ldots, \mathrm{p}$. A reta que passa por $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ e é ortogonal a $\partial \mathrm{K_p}$ é dada

por $\overrightarrow{\mathbf{r}}(\mathrm{s}) = \hat{\boldsymbol{\beta}}^{\mathrm{o}} + \mathrm{s}\,\overrightarrow{\mathbf{1}}$ (ver Figura 30).

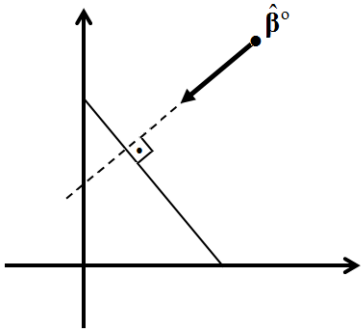


Figura 30 Estimador LASSO para o caso ortogonal.

O estimador ocorre para o primeiro valor s tal que

$$\vec{r}(s) \in \partial \mathrm{K_p} \Rightarrow \sum_{\mathrm{i}=1}^{\mathrm{p}} \left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s \right| = \mathrm{t}.$$

Se $\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} > 0$ então $\left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s \right| = \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s$, para um s pequeno. Por outro lado, se $\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} < 0$ tem-se que $\left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s \right| = -\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} + s$, e então

$$\left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s \right| = \left[\mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \right] \left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s \right|.$$

A reta $\hat{\boldsymbol{\beta}}^{\mathrm{o}} - s\,\overrightarrow{\mathbf{1}}$ atinge o bordo $\partial \mathrm{K_p}$ se

$$\sum_{\mathrm{i}=1}^{\mathrm{p}} \left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} - s \right| = \mathrm{t} \Rightarrow \sum_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \left(\left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \right| - s \right) = \mathrm{t}$$

$$\Rightarrow \sum_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \left(\left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \right| \right) - \mathrm{p}s = \mathrm{t}$$

$$\Rightarrow s = \frac{\mathrm{t} - \sum_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \left| \left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}\right)_{\mathrm{i}} \right|}{\mathrm{p}}.$$

Desta forma, o estimador LASSO é dado por

$$\left(\hat{\boldsymbol{\beta}}^{\text{lasso}}\right)_{\text{i}} \;=\; \text{sinal}\left(\hat{\boldsymbol{\beta}}^{\text{o}}\right)_{\text{i}}\left(\left|\left(\hat{\boldsymbol{\beta}}^{\text{o}}\right)_{\text{i}} - \gamma\right|\right)^{+}, \tag{4.1}$$

em que (TIBSHIRANI, 1996, p. 269)

$$\gamma = \frac{\sum_{\text{i}=1}^{p} \text{sinal}\left(\hat{\boldsymbol{\beta}}^{\text{o}}\right)_{\text{i}} \left|\left(\hat{\boldsymbol{\beta}}^{\text{o}}\right)_{\text{i}}\right|}{\text{p}}.$$

Passemos então a analisar o caso não ortogonal. Para $\text{p} = 2$, o caso não ortogonal é descrito da mesma forma que a situação ortogonal. O que ocorre agora é que a projeção na aresta do quadrado não é mais uma projeção ortogonal em relação ao produto interno usual $<,>$, mas sim em relação ao produto interno $<,>_{\text{p}}$. Para se obter um vetor ortogonal à face, observando que o vetor $(-1, 1)$ pertence à aresta, é necessário se resolver a equação

$$<(-1,1),(1,\text{a})>_{\text{p}} = (-1,1)(\text{X}'\text{X}) \begin{pmatrix} 1 \\ \text{a} \end{pmatrix} = 0.$$

A região para a qual $\beta_1^{\text{lasso}} \neq 0$ e $\beta_2^{\text{lasso}} \neq 0$ não é mais definida por retas, sendo então definida por curvas mais complexas.

Obtido o valor de a, define-se a reta $\vec{r}(s) = \hat{\boldsymbol{\beta}}^{\text{o}} + s(1, \text{a})$ e obtém-se o valor de s para o qual a reta $\vec{r}(s)$ intersecta a aresta e o raciocínio segue como no caso ortogonal. Representação desta ideia é conforme Figura 31.

De forma análoga ao estimador Ridge para o qual se pode traçar o ridge-trace para as componentes do vetor de estimativas, considerando como parâmetro de ajuste o valor t da definição do estimador LASSO, é possível traçar curvas para as componentes da estimativa LASSO. Um aspecto interessante para estas curvas é que se obtém a sequência ordenada em que as covariáveis deixam de ser zero. Desta forma, tem-se uma descrição gráfica para a seleção das covariáveis

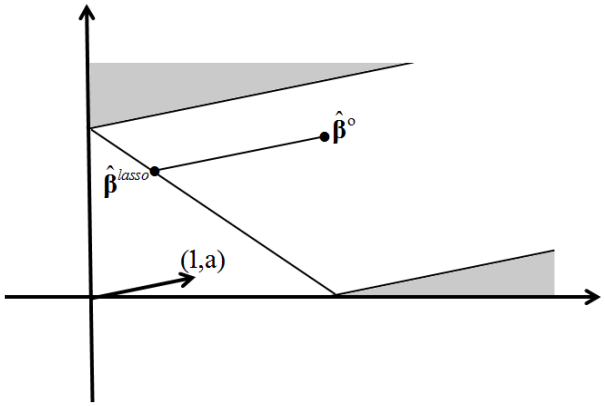


Figura 31 Estimador LASSO (caso geral).

(TIBSHIRANI, 1996, p. 271-273).

4.3 Abordagem Bayesiana aos estimadores Ridge e LASSO

O estimador ridge e o estimador LASSO podem ser obtidos facilmente a partir do paradigma bayesiano. Considere $\mathbf{y}_{\mathrm{n}\times 1}$ um vetor aleatório com densidade normal multivariada dada por

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{\mathrm{n}}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mathrm{X}\boldsymbol{\beta})'(\mathbf{y}-\mathrm{X}\boldsymbol{\beta})},$$

em que X é de dimensão $\mathrm{n} \times \mathrm{p}$ e $\boldsymbol{\beta}$ é de dimensão $\mathrm{p} \times 1$.

Supondo como priori para o vetor de parâmetros $\boldsymbol{\beta}$ uma normal multivariada da forma

$$\pi(\boldsymbol{\beta}) = \frac{(\lambda)^{\frac{\mathrm{p}}{2}}}{(2\pi)^{\frac{\mathrm{p}}{2}}} \mathrm{e}^{-\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2},$$

a posteriori é proporcional a

$$\pi(\boldsymbol{\beta}|\,\mathbf{y}) \propto \mathrm{e}^{-\frac{1}{2}\left[(\mathbf{y}-\mathrm{X}\boldsymbol{\beta})'(\mathbf{y}-\mathrm{X}\boldsymbol{\beta})-\lambda\|\boldsymbol{\beta}\|^2\right]}.$$

Uma estimativa pode então ser obtida tomando-se a moda da posteriori que ocorre

quando o expoente é minimizado, isto é,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(\text{Bayes})} &= \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left[(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})' (\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^2 \right] \right\} \\ &= \min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right\} \\ &= \hat{\boldsymbol{\beta}}^{\text{ridge}}.\end{aligned}$$

Para o LASSO, considere uma priori para o vetor $\boldsymbol{\beta}$ dada pela exponencial dupla

$$\begin{aligned}\pi(\boldsymbol{\beta}) &= \prod_{\mathrm{i}=1}^{\mathrm{p}} \frac{1}{2\tau} \mathrm{e}^{-\frac{|\beta_\mathrm{i}|}{\tau}} \\ &= \frac{1}{(2\tau)^{\mathrm{p}}} \mathrm{e}^{-\frac{1}{\tau} \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_\mathrm{i}|} \\ &= \frac{1}{(2\tau)^{\mathrm{p}}} \mathrm{e}^{-\frac{1}{\tau} \|\boldsymbol{\beta}\|_1}.\end{aligned}$$

Portanto, a posteriori é proporcional a

$$\pi(\, \boldsymbol{\beta} | \, \mathbf{y}) \propto \mathrm{e}^{-\frac{1}{2}(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) - \frac{1}{\tau} \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_\mathrm{i}|}.$$

Tomando-se a moda da posteriori tem-se

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(\text{Bayes})} &= \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})' (\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) + \frac{1}{\tau} \|\boldsymbol{\beta}\|_1 \right\} \\ &= \min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + 2\lambda \|\boldsymbol{\beta}\|_1 \right\} \\ &= \hat{\boldsymbol{\beta}}^{\text{lasso}}.\end{aligned}$$

4.4 Estimação da variância do estimador LASSO

Como o estimador LASSO é não linear e não diferenciável, é difícil obter estimativas acuradas para sua variância. De fato dois métodos foram propostos no artigo original e serão aqui discutidos em detalhes, sendo o primeiro deles utili-

zando Bootstrap.

Considere o modelo linear usual $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ com $\mathrm{E}[\boldsymbol{\varepsilon}] = 0$ e $\mathrm{var}[\boldsymbol{\varepsilon}] = \sigma^2\mathrm{I}$. Um vetor de dados $\mathbf{y}_{\mathrm{n}\times 1}$ é observado e uma estimativa $\boldsymbol{\beta}^{\mathrm{lasso}}_{\mathrm{p}\times 1}$ é obtida. Como as componentes $\mathrm{y_i}$ do vetor $\mathbf{y}$ possuem esperanças diferentes, não se pode reamostrar as componentes de $\mathbf{y}$ para se obter uma amostra bootstrap $\mathbf{y}^*$. A estratégia para se obter este novo vetor $\mathbf{y}^*$ é como se segue. Com a estimativa $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ obtida a partir do vetor de dados originais $\mathbf{y}$, obtém-se o vetor ajustado $\hat{\mathbf{y}} = \mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$. Note que $\hat{\mathbf{y}}$ não é um estimador não viesado de $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}]$. Toma-se então os resíduos $\{\varepsilon^*_{\mathrm{i}} = \hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}, \mathrm{i} = 1, \ldots, \mathrm{n}\}$. Conforme observado, não se tem $\mathrm{E}[\boldsymbol{\varepsilon}^*_{\mathrm{i}}] = 0$ e certamente $\mathrm{E}\left[(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2\right] \neq \sigma^2$. No entanto temos aqui uma analogia com o ajuste de quadrados mínimos pois se $\hat{\mathbf{y}} = \mathrm{X}\boldsymbol{\beta}^{\mathrm{ols}}$ então $\mathrm{E}[\hat{\mathbf{y}}] = \boldsymbol{\mu}$, $\mathrm{E}[(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})] = 0$ e $\frac{1}{\mathrm{n}-\mathrm{p}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2$ é um estimador não viesado de σ^2. Neste sentido é razoável considerar $\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}$ como um estimador de ε_{i}, apesar de estes resíduos não serem independentes, isto é, $\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}$ e $\hat{\mathrm{y}}_{\mathrm{j}} - \mathrm{y}_{\mathrm{j}}$ não são independentes, ao passo que ε_{i} e ε_{j} o são. Baseado nestas considerações, para $\hat{\mathbf{y}} = \mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ também será considerado $\varepsilon^*_{\mathrm{i}} = \hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}$ um estimador para o erro ε_{i}. Pode-se agora se obter amostras bootstrap da forma: $\{\varepsilon^*_1, \varepsilon^*_2, \ldots, \varepsilon^*_{\mathrm{n}}\}$ é reamostrado com repetição obtendo-se vetores $\boldsymbol{\varepsilon}^*_{\mathrm{n}\times 1}$. Com estes vetores constrói-se um vetor de pseudodados $\mathbf{y}^* = \mathbf{y} + \boldsymbol{\varepsilon}^*$, em que $\mathbf{y}$ é o vetor de dados originais. Com o vetor $\mathbf{y}^*$ obtém-se então uma nova estimativa $\hat{\boldsymbol{\beta}}^{*(\mathrm{lasso})}$. Realizando este processo várias vezes, tem-se amostras bootstrap de $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ e se pode então calcular uma estimativa para sua variância.

Uma outra abordagem é se obter uma forma fechada para uma aproximação de um estimador da variância do estimador LASSO. A ideia é utilizar uma aproximação em termos do estimador Ridge. Como $|\beta_{\mathrm{j}}| = \frac{(\beta_{\mathrm{j}})^2}{|\beta_{\mathrm{j}}|}$ tem-se que

$$\|\boldsymbol{\beta}\|_1 = \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_{\mathrm{i}}| = \sum_{\mathrm{i}=1}^{\mathrm{p}} \frac{\beta_{\mathrm{i}}^2}{|\beta_{\mathrm{i}}|}$$

e então

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{lasso}} &= \underset{\boldsymbol{\beta}}{\arg\min}\left\{\|\mathbf{y}-\mathrm{X}\boldsymbol{\beta}\|^2+\mathrm{t}\sum_{\mathrm{i}=1}^{\mathrm{p}}|\beta_{\mathrm{i}}|\right\}\\ &= \underset{\boldsymbol{\beta}}{\arg\min}\left\{\|\mathbf{y}-\mathrm{X}\boldsymbol{\beta}\|^2+\mathrm{t}\sum_{\mathrm{i}=1}^{\mathrm{p}}\frac{\beta_{\mathrm{i}}^2}{|\beta_{\mathrm{i}}|}\right\}.\end{aligned}$$

Obtido o estimador LASSO $\hat{\boldsymbol{\beta}}^{\text{lasso}}$, considere o estimador definido por

$$\hat{\boldsymbol{\beta}}^{*}=\underset{\boldsymbol{\beta}}{\arg\min}\left\{\|\mathbf{y}-\mathrm{X}\boldsymbol{\beta}\|^2+\mathrm{t}\sum_{\mathrm{i}=1}^{\mathrm{p}}\frac{(\beta_{\mathrm{i}})^2}{|(\beta^{\text{lasso}})_{\mathrm{i}}|}\right\}.$$

Ora, $\hat{\boldsymbol{\beta}}^{*}$ é simplesmente o estimador Ridge para o caso em que cada coordenada sofre um encolhimento diferente. Este caso mais geral de estimador Ridge admite uma fórmula fechada semelhante ao estimador Ridge usual dada por

$$\hat{\boldsymbol{\beta}}^{*}=\left(\mathrm{X}'\mathrm{X}+\lambda\mathrm{W}^{-}\right)^{-1}\mathrm{X}'\mathbf{y},$$

em que W^{-} é da forma da Figura 32. Se $\left(\beta^{\text{lasso}}\right)_{\mathrm{i}}=0$, então a entrada correspon-

$$\mathrm{W}^{-}=\begin{bmatrix}\frac{1}{\left|\beta_1^{\text{lasso}}\right|} & & 0\\ & \ddots & \\ 0 & & \frac{1}{\left|\beta_{\mathrm{p}}^{\text{lasso}}\right|}\end{bmatrix}$$

Figura 32 Matriz de pesos do estimador Ridge.

dente na diagonal é zero, isto é, W^{-} é a inversa generalizada da matriz diagonal. A quantidade λ é escolhida de forma que $\sum_{\mathrm{i}=1}^{\mathrm{p}}\left|\hat{\beta}_{\mathrm{i}}^{*}\right|=\mathrm{t}$, ou seja, que este estimador ridge satisfaça a mesma restrição do estimador LASSO. Tem-se então que $\operatorname{cov}\left(\hat{\boldsymbol{\beta}}^{*}\right)=(\mathrm{X}'\mathrm{X}+\lambda\mathrm{W}^{-})^{-1}\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X}+\lambda\mathrm{W}^{-})^{-1}\sigma^2$, e um estimador é então

$$\widehat{\operatorname{cov}}\left(\hat{\boldsymbol{\beta}}^{*}\right)=\left(\mathrm{X}'\mathrm{X}+\lambda\mathrm{W}^{-}\right)^{-1}\mathrm{X}'\mathrm{X}\left(\mathrm{X}'\mathrm{X}+\lambda\mathrm{W}^{-}\right)^{-1}\hat{\sigma}^2,$$

em que$\hat{\sigma}^2$ é um estimador para a variância do modelo, obtido por exemplo pelo método dos quadrados mínimos.

4.5 Erro de predição e estimação do parâmetro t

No artigo original (TIBSHIRANI, 1996) são apresentados, de forma resumida, três métodos para a estimação do parâmetro t. No sentido de completude do texto os três métodos serão explicitados, utilizando a notação do artigo original.

Considerando o modelo linear $\mathbf{y} = \mathrm{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ com $\mathrm{E}[\boldsymbol{\varepsilon}] = 0$ e $\mathrm{var}[\boldsymbol{\varepsilon}] = \sigma^2\mathrm{I}$, o erro médio (*Mean Error*) de um estimador $\hat{\boldsymbol{\beta}}$ é

$$\mathrm{ME} = \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathrm{X}\hat{\boldsymbol{\beta}} - \mathrm{X}\boldsymbol{\beta}\right\|^2\right].$$

Já o erro de predição (*Prediction Error*) será aqui considerado como

$$\mathrm{PE} = \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2\right],$$

com $\mathrm{X}\hat{\boldsymbol{\beta}}$ fixo, isto é, o modelo ajustado fica fixo, e toma-se a esperança em relação a novas observações $\mathbf{y}$. Assim,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2\right] &= \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta} + \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2\right] \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\right\|^2\right] + \mathrm{E}_{\boldsymbol{\beta}}\left[2 < \mathbf{y} - \mathrm{X}\boldsymbol{\beta}, \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}} >\right] \\
&\quad + \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2\right] \\
&= \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\boldsymbol{\varepsilon}\right\|^2\right] + 2\mathrm{E}_{\boldsymbol{\beta}}\left[< \mathbf{y} - \mathrm{X}\boldsymbol{\beta}, \mathrm{X}\boldsymbol{\beta} >\right] \\
&\quad -2\mathrm{E}\left[< \mathbf{y} - \mathrm{X}\boldsymbol{\beta}, \mathrm{X}\hat{\boldsymbol{\beta}} >\right] + \mathrm{E}_{\boldsymbol{\beta}}\left[\left\|\mathrm{X}\boldsymbol{\beta} - \mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2\right] \\
&= \mathrm{n}\sigma^2 + 2 < \mathrm{E}\left[\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\right], \mathrm{X}\boldsymbol{\beta} > -2\mathrm{E}\left[< \mathbf{y} - \mathrm{X}\boldsymbol{\beta}, \mathrm{X}\hat{\boldsymbol{\beta}} >\right] \\
&\quad + \mathrm{ME}
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{n}\sigma^2 + 0 - 2\mathrm{E}\left[< \mathbf{y} - \mathrm{X}\boldsymbol{\beta}, \mathrm{X}\hat{\boldsymbol{\beta}} >\right] + \mathrm{ME} \\
&= \mathrm{n}\sigma^2 - 2\left[< \mathrm{E}\left[\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\right], \mathrm{X}\hat{\boldsymbol{\beta}} >\right] + \mathrm{ME} \\
&= \mathrm{n}\sigma^2 + \mathrm{ME}.
\end{aligned}
$$

O erro de predição do estimador LASSO pode ser estimado por exemplo com o método *fivefold cross-validation*. Neste método os dados são divididos aleatoriamente em cinco partes iguais. Em quatro delas simultaneamente o modelo é ajustado e a outra restante é utilizada como conjunto de teste do modelo ajustado. O processo é então repetido cinco vezes e em seguida toma-se a média do erro de predição. No caso do LASSO o vetor $\mathbf{y}$ é particionado em cinco partes iguais. Uma delas é escolhida. Retira-se então da matriz X as linhas correspondentes a esta parte escolhida. Com esta matriz reduzida e o vetor de dados reduzido é obtida uma estimativa LASSO $\hat{\boldsymbol{\beta}}_1$. O processo é repetido cinco vezes obtendo assim $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_5$. Observa-se que apesar destes vetores terem sido obtidos pela matriz X reduzida de um quinto das linhas, os vetores $\hat{\boldsymbol{\beta}}_{\mathrm{i}}$ obtidos são vetores em $\mathbb{R}^{\mathrm{p}}$. Para o erro de predição a matriz original X é aplicada nos vetores $\hat{\boldsymbol{\beta}}_{\mathrm{i}}$ obtendo-se vetores ajustados $\hat{\mathbf{y}}_{\mathrm{i}}$ em $\mathbb{R}^{\mathrm{n}}$. Calcula-se então, para as componentes ao um quinto das coordenadas do vetor de dados originais o quadrado da diferença do valor ajustado e dos valores observados. O processo é repetido cinco vezes e toma-se a média destes. A escolha do valor do parâmetro t é obtida fazendo a reparametrização $\mathrm{s} = \dfrac{\mathrm{t}}{\sum\limits_{\mathrm{i}=1}^{\mathrm{p}} \left|\beta_{\mathrm{i}}^{\mathrm{ols}}\right|}$. Como o valor de t na estimação LASSO só é interessante para $0 < \mathrm{t} \leq \sum\limits_{\mathrm{i}=1}^{\mathrm{p}} \left|\beta_{\mathrm{i}}^{\mathrm{ols}}\right|$, o novo parâmetro s varia de zero a um. Faz-se então uma partição de $[0, 1]$ da forma $\mathrm{s} = 0, \frac{1}{\ell}, \frac{2}{\ell}, \ldots, \frac{\ell-1}{\ell}, 1$ para ℓ suficientemente grande e calcula-se a estimativa do erro de predição para os estimadores LASSO para cada um dos valores de s. Toma-se então por estimativa de t o valor de s correspondente que gera o menor erro de predição. Observa-se que como o erro de predição PE

e o erro quadrático médio ME diferem pelo valor da variância σ^2, a computação pode ser feita em relação a ME. Tem-se então uma fórmula computacionalmente mais simples, uma vez que para modelos lineares se tem

$$\mathrm{ME} = \mathrm{E}\left[\left\|\mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} - \mathrm{X}\boldsymbol{\beta}\right\|^2\right] = \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} - \boldsymbol{\beta}\right)'\mathrm{X}'\mathrm{X}\left(\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} - \boldsymbol{\beta}\right)\right],$$

e portanto pode ser estimado por $\frac{1}{5}\sum_{\mathrm{i}=1}^{5}\left(\hat{\boldsymbol{\beta}}_{\mathrm{i}}^{\mathrm{lasso}} - \boldsymbol{\beta}\right)'\mathrm{X}'\mathrm{X}\left(\hat{\boldsymbol{\beta}}_{\mathrm{i}}^{\mathrm{lasso}} - \boldsymbol{\beta}\right)$.

O segundo método é baseado na aproximação do estimador $\boldsymbol{\beta}^{\mathrm{lasso}}$ pelo estimador Ridge $\boldsymbol{\beta}^*(\lambda) = (\mathrm{X}'\mathrm{X} + \lambda\mathrm{W}^-)^{-1}\mathrm{X}'\mathbf{y}$, para $\mathrm{W} = \mathrm{diag}\left(\left|\beta_{\mathrm{j}}^{\mathrm{lasso}}\right|\right)$, conforme descrito em 4.4. O erro de predição do estimador LASSO $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ será aproximado pelo erro de predição do estimador $\boldsymbol{\beta}^*(\lambda)$ que, por ser um estimador linear, é calculado por um processo de *cross-validation* tipo *leave one out*. Seja $\boldsymbol{\beta}^{*(\mathrm{k})}(\lambda)$ o estimador obtido omitindo-se a k-ésima coordenada do vetor de dados $\mathbf{y}$ e a linha correspondente na matriz $(\mathrm{X}'\mathrm{X} + \lambda\mathrm{W}^-)^{-1}\mathrm{X}'$. Obtém-se então a estimativa do erro de predição

$$\mathrm{P}(\lambda) = \frac{1}{\mathrm{n}}\left\|\mathrm{X}\boldsymbol{\beta}^{*(\mathrm{k})}(\lambda) - \mathbf{y}\right\|^2 = \frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(\left(\mathrm{X}\boldsymbol{\beta}^{*(\mathrm{k})}(\lambda)\right)_{(\mathrm{k})} - \mathrm{y}_{\mathrm{k}}\right)^2,$$

denominada na literatura de PRESS, acrônimo de *The Predicted Residual Sum of Squares*. De forma surpreendente Golub; Heath; Wahba (1979) demonstram que $\mathrm{P}(\lambda)$ admite a seguinte expressão matricial simples

$$\mathrm{P}(\lambda) = \frac{1}{\mathrm{n}}\left\|\mathrm{D}(\lambda)\left(\mathrm{I} - \mathrm{X}\left(\mathrm{X}'\mathrm{X} + \lambda\mathrm{W}^-\right)^{-1}\mathrm{X}'\mathbf{y}\right)\right\|^2,$$

em que $\mathrm{D}(\lambda) = \mathrm{diag}\left(\frac{1}{1-\mathrm{a}_{\mathrm{ii}}}\right)$ e $\mathrm{a}_{\mathrm{ii}} = \left(\mathrm{X}(\mathrm{X}'\mathrm{X} + \lambda\mathrm{W}^-)^{-1}\mathrm{X}'\right)_{\mathrm{ii}}$.

A utilização desta fórmula permite obter para o caso do estimador Ridge o valor ótimo para o parâmetro λ simplesmente obtendo o valor de λ que minimiza $\mathrm{P}(\lambda)$, isto é, que minimiza o erro de predição. Golub, op. cit., apresentam uma generalização deste método, que consiste em uma versão invariante por rotação.

Aplica-se ao vetor de dados $\mathbf{y}$ uma matriz especial obtendo-se uma nova estimativa para o erro de predição dada por

$$\mathrm{V}(\lambda)=\frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(\left(\mathrm{X}\boldsymbol{\beta}^{*(\mathrm{k})}(\lambda)\right)_{\mathrm{k}}-\mathrm{y}_{\mathrm{k}}\right)^{2}\omega_{\mathrm{k}}(\lambda),$$

em que

$$\omega_{\mathrm{k}}(\lambda)=\frac{1-\mathrm{a}_{\mathrm{kk}}(\lambda)}{1-\frac{1}{\mathrm{n}}\mathrm{tr}\left(\mathrm{A}(\lambda)\right)}.$$

A estimativa para o parâmetro t é obtida como um minimizador de $\mathrm{V}(\lambda)$ (t e λ estão univocamente relacionados). Este procedimento recebeu o nome de GCV (Generalized Cross Validation). No artigo original do LASSO, é proposta uma estatística semelhante da forma

$$\mathrm{GCV}(\mathrm{t})=\frac{1}{\mathrm{n}}\frac{\left\|\mathrm{X}\boldsymbol{\beta}^{\mathrm{lasso}}(\mathrm{t})-\mathbf{y}\right\|^{2}}{\left\{1-\frac{1}{\mathrm{n}}\mathrm{tr}\left(\mathrm{X}(\mathrm{X}'\mathrm{X}+\lambda\mathrm{W}^{-})^{-1}\mathrm{X}'\right)\right\}^{2}}.$$

O terceiro método é baseado no estimador não viesado de Stein (SURE) (ver Seção 7.2). Suponha $\mathrm{Z}\sim\mathrm{N}_{\mathrm{n}}(\boldsymbol{\mu},\mathrm{I})$. Se $\hat{\boldsymbol{\mu}}$ é um estimador de $\boldsymbol{\mu}$, fazendo $\hat{\boldsymbol{\mu}}=\mathbf{z}+\mathrm{g}(\mathbf{z})$, então

$$\mathrm{E}_{\boldsymbol{\mu}}\left[\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^{2}\right]=\mathrm{n}+\mathrm{E}_{\boldsymbol{\mu}}\left[\|\mathrm{g}(\mathbf{z})\|^{2}+2\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\partial\mathrm{g}_{\mathrm{i}}}{\partial\mathrm{z}_{\mathrm{i}}}\right],$$

e pelo método dos momentos tem-se o estimador não viesado para o risco $\mathrm{E}_{\boldsymbol{\mu}}\left[\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^{2}\right]$ do estimador $\hat{\boldsymbol{\mu}}$, dado por $\mathrm{n}+\|\mathrm{g}(\mathbf{z})\|^{2}+2\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\partial\mathrm{g}_{\mathrm{i}}}{\partial\mathrm{z}_{\mathrm{i}}}$. Supondo ortogonalidade, isto é, $\mathrm{X}'\mathrm{X}=\mathrm{I}$, tem-se uma fórmula explícita para o estimador LASSO,

$$\hat{\beta}_{\mathrm{j}}^{\mathrm{lasso}}=\mathrm{sinal}\left(\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right)\left(\left|\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right|-\gamma\right)^{+},$$

em que γ é definido por uma relação com $\mathrm{t}=\sum_{\mathrm{i}=1}^{\mathrm{p}}\left|\hat{\beta}_{\mathrm{j}}^{\mathrm{lasso}}\right|$. Estimando o erro padrão de $\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}$ por $\hat{\tau}=\frac{\hat{\sigma}}{\sqrt{\mathrm{n}}}$ em que $\hat{\sigma}^{2}=\frac{1}{\mathrm{n}-\mathrm{p}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_{\mathrm{i}}-\hat{\mathrm{y}}_{\mathrm{i}})^{2}$, então $\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}$, $\mathrm{j}=1,\ldots,\mathrm{n}$,

são aproximadamente independentes e normais padrão. Em razão da natureza geométrica do processo de estimação LASSO, pode-se afirmar que o estimador $\hat{\beta}_{\mathrm{j}}^{\mathrm{lasso}} = \mathrm{sinal}\left(\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right)\left(\left|\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right| - \gamma\right)^{+}$ é próximo ao estimador definido por

$$\hat{\boldsymbol{\mu}}_{\mathrm{i}} = \mathrm{sinal}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right)\left(\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| - \gamma\right)^{+},$$

conforme Figura 33. Fazendo $\mathrm{z}_{\mathrm{i}} = \frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}$ tem-se

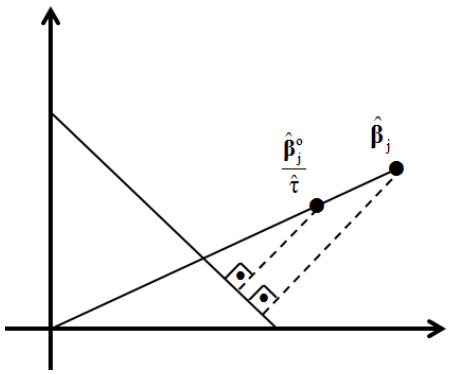


Figura 33 Aproximação do estimador LASSO.

$$\hat{\boldsymbol{\mu}}_{\mathrm{i}} = \mathrm{sinal}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right)\left(\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| - \gamma\right)^{+} = \frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}} - \frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}} + \mathrm{sinal}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right)\left(\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| - \gamma\right)^{+}.$$

Definindo a função $\mathrm{g}(\mathbf{z}) = (\mathrm{g}_1(\mathrm{z}_1), \ldots, \mathrm{g}_{\mathrm{p}}(\mathrm{z}_{\mathrm{p}}))$ por

$$\mathrm{g}_{\mathrm{i}}(\mathrm{z}_{\mathrm{i}}) = \mathrm{g}_{\mathrm{i}}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right) = -\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}} + \mathrm{sinal}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right)\left(\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| - \gamma\right)^{+},$$

se $\hat{\beta}_{\mathrm{j}}^{\mathrm{o}} < 0$ e $\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| < \gamma$ então $\mathrm{g}_{\mathrm{i}}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right) = -\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}} - 0.$

Se $\hat{\beta}_{\mathrm{j}}^{\mathrm{o}} < 0$ e $\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| > \gamma$ então $\mathrm{g}_{\mathrm{i}}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right) = -\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}} - \left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| + \gamma = \gamma.$

Se $\hat{\beta}_{\mathrm{j}}^{\mathrm{o}} > 0$ e $\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| < \gamma$ então $\mathrm{g}_{\mathrm{i}}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right) = -\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}.$

Finalmente, para $\hat{\beta}_{\mathrm{j}}^{\mathrm{o}} > 0$ e $\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| > \gamma$ a função g é $\mathrm{g_i}\left(\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right) = -\gamma$. Note que $\frac{\partial \mathrm{g_i}}{\partial \frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}} = 1$ sempre que $\left|\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right| < \gamma$.

Em suma, a função $\hat{\boldsymbol{\mu}} = \mathbf{z} + \mathrm{g}(\mathbf{z})$ para $\mathrm{g}(\mathbf{z}) = -\,\mathrm{z} + \mathrm{sinal}(\mathrm{z})(|\mathrm{z}| - \gamma)^{+}$ e $\mathrm{z} = \frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}$ possui coordenadas com gráfico dado pela Figura 34. Portanto, o estimador não viesado de Stein para o erro quadrático médio, $\widehat{\mathrm{E}_{\boldsymbol{\mu}}}\left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\right] = \mathrm{p} + \|\mathrm{g}(\mathbf{z})\|^2 + 2\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial \mathrm{g_i}}{\partial \mathrm{z_i}}$, tem sua expressão dada por

$$\mathrm{R}\left\{\hat{\boldsymbol{\beta}}(\gamma)\right\} \approx \hat{\tau}^2 \left\{ \mathrm{p} - 2\#\left(\mathrm{j}; \left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right| < \gamma\right) + \sum_{\mathrm{j}=1}^{\mathrm{p}} \max\left(\left|\frac{\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}}{\hat{\tau}}\right|, \gamma\right)^2 \right\}.$$

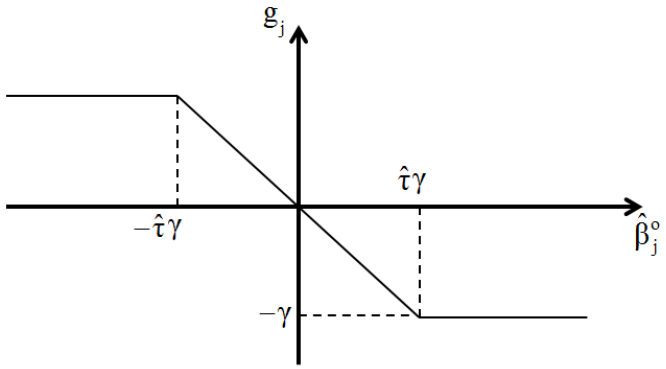


Figura 34 Gráfico das funções coordenadas.

A equação que relaciona os parâmetros t e γ pode ser explicitada a partir do fato de que $\hat{\beta}_{\mathrm{j}}^{\mathrm{o}} = \mathrm{sinal}\left(\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right)\left(\left|\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right| - \gamma\right)^{+}$ e, uma vez que $\sum_{\mathrm{j}=1}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right| = \mathrm{t}$, tem-se então

$$\sum_{\mathrm{j}=1}^{\mathrm{p}} \left[\left|\mathrm{sinal}\left(\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right)\right|\left(\left(\left|\hat{\beta}_{\mathrm{j}}^{\mathrm{o}}\right| - \gamma\right)^{+}\right)\right] = \mathrm{t},$$

$$\Rightarrow \hat{\mathrm{t}} = \sum_{\mathrm{j}=1}^{\mathrm{p}} \left(\hat{\beta}_{\mathrm{j}}^{\mathrm{o}} - \hat{\gamma}\right)^{+}.$$

O artigo apresenta uma justificativa heurística para a razão de que o caso ortogo-

nal também pode ser aplicado como uma aproximação para o risco do estimador LASSO no caso geral.

4.6 O caso $\mathrm{p} > \mathrm{n}$

É usual em problemas de regressão, principalmente em aplicações genéticas, que o número de covariáveis p seja maior do que o número de observações n. É observado sem demonstração em Zou; Hastie (2005) que o método LASSO seleciona no máximo n covariáveis, sendo esta uma limitação do método que foi superada com o método Elasticnet. Na literatura não foi encontrada uma demonstração explícita deste fato. Segue portanto uma demonstração baseada em argumentos geométricos que justifica tal fato.

No caso $\mathrm{p} > \mathrm{n}$, a matriz $\mathrm{X}_{\mathrm{n}\times\mathrm{p}}$ não é injetiva e possui kernel, isto é, o subespaço $\mathrm{Ker(X)} = \{\boldsymbol{\beta},\ \mathrm{X}\boldsymbol{\beta} = \mathbf{0}\}$. O estimador LASSO é definido então por

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\mathrm{p}}}{\arg\min} \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

$\hat{\boldsymbol{\beta}}^{lasso}$ é obtido geometricamente projetando-se o estimador de quadrados mínimos, pela métrica de Mahalanobis, no convexo definido por $\|\boldsymbol{\beta}\|_1 \leq \mathrm{t}$. Neste caso, para se obter o estimador de quadrados mínimos faz-se necessário a utilização de uma inversa generalizada de $\mathrm{X}'\mathrm{X}$, definindo a estimativa dada por $\hat{\tilde{\boldsymbol{\beta}}}^{\mathbf{ols}} = (\mathrm{X}'\mathrm{X})^{-}\mathrm{X}'\mathbf{y}$. Neste caso, tem-se que se $\boldsymbol{\beta} \in Ker(\mathrm{X})$ então

$$\begin{aligned}\left\|\mathbf{y} - \mathrm{X}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}} + \boldsymbol{\beta}\right)\right\|^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{\mathrm{ols}} + \boldsymbol{\beta}\right\|_1 &= \left\|\mathbf{y} - \mathrm{X}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)\right\|^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{\mathrm{ols}} + \boldsymbol{\beta}\right\|_1 \\ &\leq \left\|\mathbf{y} - \mathrm{X}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)\right\|^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right\|_1 + \lambda\|\boldsymbol{\beta}\|_1.\end{aligned}$$

Tem-se então que se deve minimizar $\|\boldsymbol{\beta}\|_1$ para $\boldsymbol{\beta} \in \mathrm{Ker(X)}$. Para a norma $\|\ \|_1$ isto ocorre quando o subespaço paralelo ao $\mathrm{Ker(X)}$, e que contém o $\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$, intersec-

ciona os planos coordenados, conforme Figura 35 . Supondo sem perda de generalidade que a intersecção seja transversal, então se tem no máximo uma intersecção com o plano coordenado de dimensão n, e portanto no máximo n covariáveis são selecionadas.

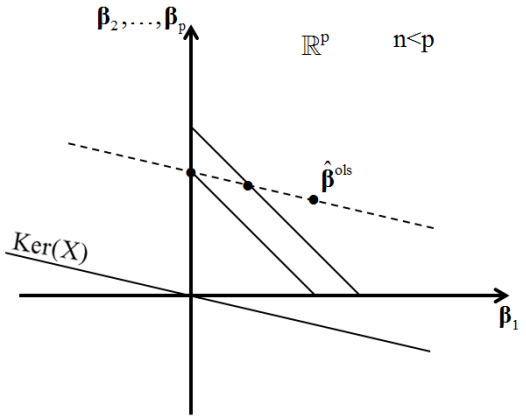


Figura 35 $\mathrm{Ker}\,(\mathrm{X})$ e o convexo $\mathrm{K_p}$

4.7 GARROTE

O método foi proposto originalmente no artigo clássico *Better Subset Regression Using the Nonnegative Garrote* (BREIMAN, 1995), como uma forma seleção de covariáveis em regressão linear. Este método foi que motivou a definição do estimador LASSO. Neste trabalho, o método GARROTE será discutido de forma simplificada. No problema de regressão linear $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, com $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ sendo o estimador de quadrados mínimos, considere o conjunto

$$\mathrm{C_p} = \left\{ (\mathrm{c_1}, \ldots, \mathrm{c_p});\, \mathrm{c_j} \geq 0 \ \mathrm{e} \sum_{\mathrm{j=1}}^{\mathrm{p}} \mathrm{c_j} \leq \mathrm{t} \right\}.$$

Para $\mathrm{p} = 2$ e $\mathrm{p} = 3$ estes conjuntos são, respectivamente, um triângulo e um tetraedro, conforme Figura 36. A partir do convexo $\mathrm{C_p}$, é construído um convexo $\mathrm{K_p}$ no

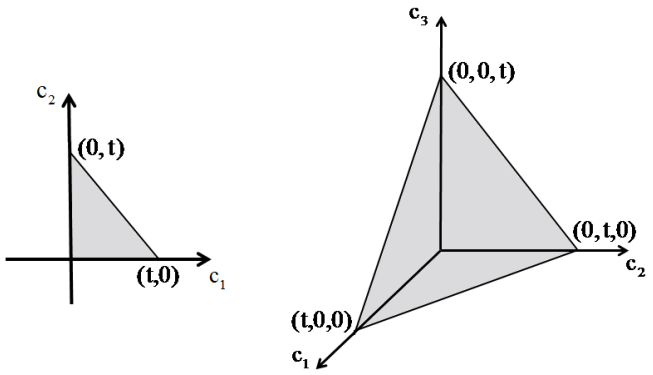


Figura 36 O convexo $\mathrm{C_p}$ em $\mathbb{R}^2$ e $\mathbb{R}^3$

espaço paramétrico da forma como se segue. Seja $\hat{\boldsymbol{\beta}}^{\mathrm{o}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_{\mathrm{p}}\right)$ a estimativa de quadrados mínimos. Considere a transformação linear $\mathrm{A} : \mathbb{R}^{\mathrm{p}} \to \mathbb{R}^{\mathrm{p}}$ em que A é a matriz diagonal $\mathrm{A} = \mathrm{diag}(\hat{\boldsymbol{\beta}}_{\mathrm{i}}^{\mathrm{o}})$. Logo, a imagem de $\mathrm{C_p}$ por A, $\mathrm{K_p} = \mathrm{A}\left(\mathrm{C_p}\right)$, é

$$\mathrm{K_p} = \left\{ \left(\mathrm{c}_1\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}, \ldots, \mathrm{c_p}\hat{\boldsymbol{\beta}}_{\mathrm{p}}^{\mathrm{o}}\right), \mathrm{c_j} \geq 0, \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{c_j} \leq \mathrm{t} \right\}.$$

A forma de $\mathrm{K_p}$ é similar à de $\mathrm{C_p}$, possuindo entretanto inclinações diferentes. Para $\mathrm{p} = 2$, tem-se a representação na Figura 37. É possível observar que se $\mathrm{t} = \mathrm{p}$ então $(1, \ldots, 1) \in \mathrm{C_p}$ e consequentemente a estimativa de quadrados mínimos $\left(\hat{\beta}_1, \ldots, \hat{\beta}_{\mathrm{p}}\right) \in \mathrm{K_p}$. Se $\mathrm{t} < \mathrm{p}$ então $\hat{\boldsymbol{\beta}}^{\mathrm{o}} \notin \mathrm{K_p}$. O método de estimação GARROTE consiste em se projetar ortogonalmente, em relação à métrica $< \, , \, >_{\mathrm{p}}$, $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ em $\mathrm{K_p}$. No espaço de dados, a projeção de $\mathbf{y}$ é ortogonal em relação ao produto usual em $\mathrm{K} = \mathrm{X}(\mathrm{K_p})$.

Analiticamente, o GARROTE pode ser definido como

$$\hat{\boldsymbol{\beta}}^{\mathrm{garrote}} = \arg\min \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathbf{y}_{\mathrm{i}} - \sum_{\mathrm{j}=1}^{\mathrm{p}} x_{\mathrm{ij}} \mathrm{c}_{\mathrm{j}} \hat{\boldsymbol{\beta}}_{\mathrm{j}}^{\mathrm{o}} \right)^2$$

sujeito à restrição $\sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{c}_{\mathrm{j}} \leq \mathrm{t}$, $\mathrm{c}_{\mathrm{j}} \geq 0$ (TIBSHIRANI, 1996; BREIMAN, 1995). A projeção no processo GARROTE possui tendência a ocorrer nas faces de menor dimensão e portanto tende a obter o estimador GARROTE com vários componentes nulos, o que pode ser visto como um método de seleção de variáveis. Como se pode observar, o método LASSO é semelhante ao GARROTE, diferindo no fato de que o convexo K_{p} no GARROTE depende da estimativa de quadrados mínimos.

Um caso particular é quando se tem $\mathrm{t} \geq \mathrm{p}$. Nesta situação, $\sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{c}_{\mathrm{j}} = \mathrm{p} \leq \mathrm{t}$ é satisfeito por $\mathrm{c}_1 = \mathrm{c}_2 = \ldots = \mathrm{c}_{\mathrm{p}} = 1$, implicando que $\hat{\boldsymbol{\beta}}^{\mathrm{o}} \in \mathrm{K}_{\mathrm{p}}$ e então o estimador GARROTE coincide com o estimador de quadrados mínimos $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$.

Como exemplo seja o caso ortogonal, em que $\mathrm{X}'\mathrm{X} = \mathrm{I}$ e $\mathrm{p} = 2$ (Figura 37). O vetor normal à reta aresta que passa por $(0, \mathrm{t}\,\hat{\boldsymbol{\beta}}_2^{\mathrm{o}})$ e $(\mathrm{t}\,\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}, 0)$ é

$$< (\mathrm{t}\,\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}, -\mathrm{t}\,\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}), (1, \mathrm{a}) >= 0$$

$$\mathrm{t}\,\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - \mathrm{a}\,\mathrm{t}\,\hat{\boldsymbol{\beta}}_2^{\mathrm{o}} = 0 \Rightarrow \mathrm{a} = \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}.$$

A reta que passa por $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ e tem este vetor como vetor diretor é

$$\mathrm{r}(s) = \hat{\boldsymbol{\beta}}^{\mathrm{o}} + s \left(1, \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}} \right).$$

A intersecção desta reta com a reta aresta ocorre para

$$\frac{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}} + s \dfrac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}}{\mathrm{t}\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - (\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + s)} = \frac{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}$$

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} + s\left(\frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}\right)^2 = \mathrm{t}\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - s$$

$$s\left(\left(\frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}\right)^2 + 1\right) = t\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - 2\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}$$

$$s = \frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}} - 2\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{1 + \left(\frac{\hat{\boldsymbol{\beta}}_1^{\mathrm{o}}}{\hat{\boldsymbol{\beta}}_2^{\mathrm{o}}}\right)^2}.$$

Obtendo os estimadores:

$$\left(\left(1 + \frac{\mathrm{t} - 2}{1 + \left(\frac{\hat{\beta}_1^{\mathrm{o}}}{\hat{\beta}_2^{\mathrm{o}}}\right)^2}\right)^+ \hat{\beta}_1^{\mathrm{o}}\ ,\ \left(1 + \frac{\mathrm{t} - 2}{1 + \left(\frac{\hat{\beta}_2^{\mathrm{o}}}{\hat{\beta}_1^{\mathrm{o}}}\right)^2}\right)^+ \hat{\beta}_2^{\mathrm{o}}\right).$$

A fórmula anterior difere da fórmula apresentada em Tibshirani (1996, p. 269).

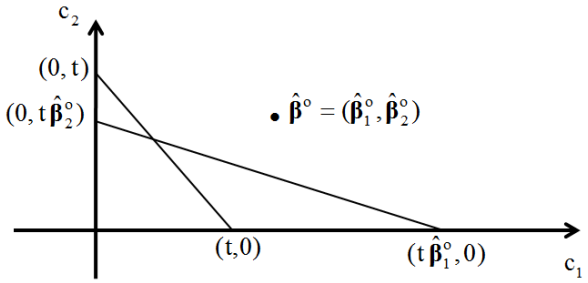


Figura 37 Penalização do método GARROTE.

O grande problema relativo ao método GARROTE é que como este utiliza a estimativa de quadrados mínimos na construção do convexo $\mathrm{K_p}$, se o estimador de quadrados mínimos apresenta grande variabilidade, o mesmo ocorrerá com o estimador GARROTE.

5 O método de regressão *Elastic Net*

Este método de regressão pretende obter o melhor entre os dois métodos de regressão, Ridge e LASSO, porém com novas propriedades que estes métodos carecem. A ideia é bastante simples, consistindo em minimizar a soma de quadrados do resíduo restrito a uma combinação linear das restrições do método Ridge e do LASSO.

Para uma regressão múltipla $\mathbf{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, considerando o modelo centrado e as covariáveis na forma de correlação, isto é, $\sum \mathrm{y_i} = 0$, $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_{ij}} = 0$, $\sum_{\mathrm{i}=1}^{\mathrm{n}} {\mathrm{x_{ij}}}^2 = 1$ com $\mathrm{j} = 1, \ldots, \mathrm{p}$, a penalização da soma de quadrados será dada por

$$\lambda_1 \|\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|_1,$$

em que $\|\boldsymbol{\beta}\|^2 = \sum_{\mathrm{j}=1}^{\mathrm{p}} \boldsymbol{\beta}_{\mathrm{j}}^2$ e $\|\boldsymbol{\beta}\|_1 = \sum_{\mathrm{j}=1}^{\mathrm{p}} \left|\boldsymbol{\beta}_j\right|$. Portanto, a função a ser minimizada é dada por $\mathrm{L}\left(\boldsymbol{\beta}, \lambda_1, \lambda_2\right) = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|_1$. O estimador obtido será denominado estimador *elastic net* ingênuo, representado por

$$\hat{\boldsymbol{\beta}}_{eni} = \arg\min_{\beta} \mathrm{L}\left(\beta, \lambda_1, \lambda_2\right)$$

5.1 O estimador *elastic net*

Em termos de soma de quadrados restrita (penalizada) minimizar a função $\mathrm{L}\left(\boldsymbol{\beta}, \lambda_1, \lambda_2\right)$ é equivalente a minimizar $\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|$ sujeito à restrição

$$\begin{aligned}
\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|^2 \leq t \quad &\Rightarrow \quad \frac{\lambda_1}{\lambda_1 + \lambda_2} \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \|\boldsymbol{\beta}\|^2 \leq \frac{t}{\lambda_1 + \lambda_2} \\
&\Rightarrow \quad \left(1 - \frac{\lambda_2}{\lambda_1 + \lambda_2}\right) \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \|\boldsymbol{\beta}\|^2 \leq \frac{t}{\lambda_1 + \lambda_2} \\
&\Rightarrow \quad (1 - \alpha) \|\boldsymbol{\beta}\|_1 + \alpha \|\boldsymbol{\beta}\|^2 \leq t'.
\end{aligned}$$

$(1-\alpha)\left\|\boldsymbol{\beta}\right\|_1 + \alpha\|\boldsymbol{\beta}\|^2$ é denominada de penalidade *elastic net* e é uma combinação convexa entre as penalidades que definem a estimação Ridge e LASSO. Pode-se observar que para $\alpha = 1$ se tem a estimação Ridge e se $\alpha = 0$ a estimação resultante é a do LASSO.

O conjunto $\left\|\boldsymbol{\beta}\right\|_1 \leq t$ é convexo, mas não estritamente convexo. Já o conjunto $\|\boldsymbol{\beta}\|^2 \leq t$ é estritamente convexo, e portanto para $\alpha \neq 0$, o conjunto definido por $(1-\alpha)\left\|\boldsymbol{\beta}\right\|_1 + \alpha\|\boldsymbol{\beta}\|^2$ é estritamente convexo. Logo o modelo $\mathrm{K_p}$ é o convexo

$$\mathrm{K_p} = \left\{\boldsymbol{\beta}, (1-\alpha)\left\|\boldsymbol{\beta}\right\|_1 + \alpha\|\boldsymbol{\beta}\|^2 \leq \mathrm{t}\right\}.$$

Para analisar a forma deste conjunto convexo basta analisar o seu bordo, isto é, analisar a igualdade $(1-\alpha)\left\|\boldsymbol{\beta}\right\|_1 + \alpha\|\boldsymbol{\beta}\|^2 = \mathrm{t}$. Como exemplo, para $\mathrm{t} = 2$, $\alpha = 0,5$ e $p = 2$ se tem $(0,5)\left\|\boldsymbol{\beta}\right\|_1 + (0,5)\|\boldsymbol{\beta}\|^2 = 2$. Assim, se $\left\|\boldsymbol{\beta}\right\|_1 = 2$ e $\|\boldsymbol{\beta}\|^2 = 2$, tem-se a representação conforme Figura38.

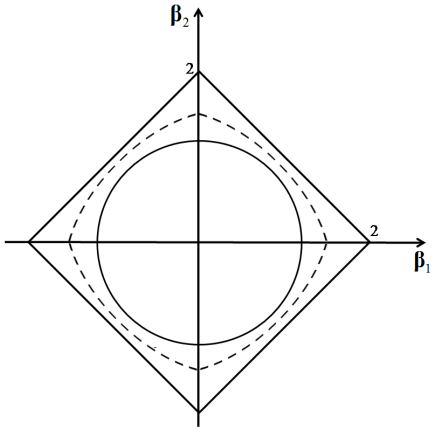


Figura 38 $\mathrm{K_p}$ na estimação *elastic net*.

Novamente o processo de estimação *elastic net* consiste em projetar com distância mínima o estimador de quadrados mínimos $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ em $\mathrm{K_p}$. Geometricamente a projeção é obtida como o ponto de tangência entre a família de elipses centrada

em $\hat{\boldsymbol{\beta}}^{\mathrm{o}}$ e a curva $(1-\alpha)\,\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|^2 = \mathrm{t}$, conforme representação na Figura 39.

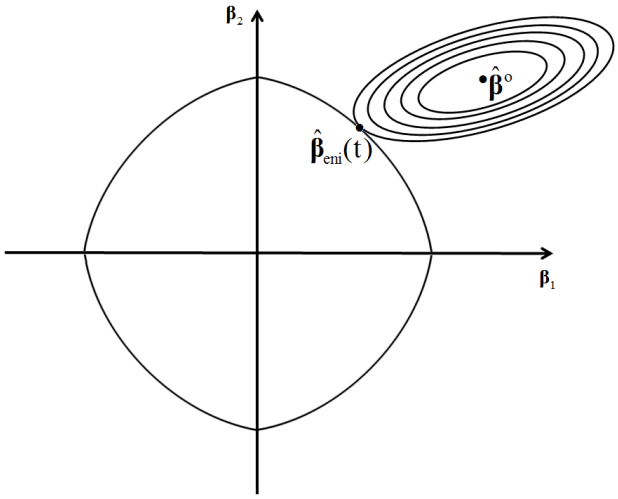


Figura 39 Geometria do estimador *elastic net*.

Observando novamente a Figura 38 tem-se que nos pontos $(\pm\,\mathrm{t}\,,0)$ e $(0\,,\pm\,\mathrm{t})$ a curva não é diferenciável, isto é, se tem "pontas". Esta é a razão pela qual o método *elasticnet* mantém as propriedades do LASSO de seleção de covariáveis.

No espaço de dados o método pode ser visualizado conforme descrito a seguir. A aplicação X leva o conjunto convexo $\mathrm{K_p}$, $(1-\alpha)\,\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|^2 = \mathrm{t}$, em um conjunto convexo $\mathrm{K} \subset \mathrm{Im(X)}$. O processo é simplesmente obter o vetor $\mathrm{X}\boldsymbol{\beta}$ em K que esteja o mais próximo de $\mathbf{y}$ (Figura 40).

O processo do *elastic net* ingênuo pode ser visto como um processo de duas etapas: uma regressão Ridge e uma regressão tipo LASSO. Este fato será fundamental na construção de algoritmos eficientes para o cálculo de estimativas *elasticnet*. A ideia é que a regressão Ridge pode ser obtida com o emprego de estimadores mistos (GRUBER, 1998; COSTA, 2015), isto é, utilizando um modelo

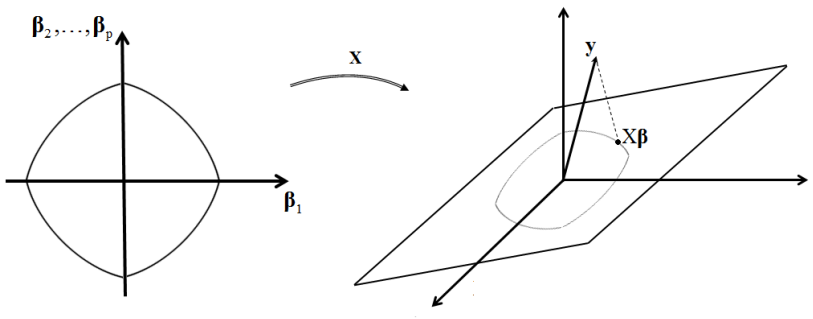


Figura 40 O método de regressão *elastic net*.

linear aumentado a partir do modelo original. Este procedimento está descrito a seguir.

Primeiramente faz-se uma reparametrização, em que os novos parâmetros são da forma $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. A partir do conjunto de dados $(\mathbf{y}_{\mathrm{n}\times 1}, \mathrm{X}_{\mathrm{n}\times \mathrm{p}})$ define-se um novo conjunto de dados (dados aumentados) que podem por exemplo ser provenientes de um outro experimento anteriormente realizado. Com estes dados aumentados a regressão fica da forma $\left(\mathbf{y}^*_{(\mathrm{n}+\mathrm{p})\times 1}, \mathrm{X}^*_{(\mathrm{n}+\mathrm{p})\times \mathrm{p}}\right)$, em que $\mathbf{y}^*_{(\mathrm{n}+\mathrm{p})\times 1} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$ e $\mathrm{X}^*_{(\mathrm{n}+\mathrm{p})\times \mathrm{p}} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathrm{X}_{\mathrm{n}\times \mathrm{p}} \\ \sqrt{\lambda_2}\mathrm{I}_{\mathrm{p}\times \mathrm{p}} \end{pmatrix}$. Assim, tem-se então uma nova regressão em relação aos novos parâmetros $\boldsymbol{\beta}^*$ dada por $\mathbf{y}^* = \mathrm{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$. O estimador de quadrados mínimos desta regressão é

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{*\,\mathrm{o}} &= \left(\mathrm{X}^{*\prime}\mathrm{X}^*\right)^{-1}\mathrm{X}^{*\prime}\mathbf{y}^* \\
&= \frac{1+\lambda_2}{\sqrt{1+\lambda_2}} \left(\left(\mathrm{X}', \sqrt{\lambda_2}\mathrm{I}\right) \begin{pmatrix} \mathrm{X} \\ \sqrt{\lambda_2}\mathrm{I} \end{pmatrix} \right)^{-1} \left(\mathrm{X}', \sqrt{\lambda_2}\mathrm{I}\right) \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \\
&= \sqrt{1+\lambda_2}\left(\mathrm{X}'\mathrm{X} + \lambda_2\mathrm{I}\right)^{-1}\mathrm{X}'\mathbf{y}.
\end{aligned}$$

Logo, como $\hat{\boldsymbol{\beta}}^{*\,\mathrm{o}} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}$, tem-se que $\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{X}+\lambda_2\mathrm{I})^{-1}\mathrm{X}'\mathbf{y} = \hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda_2)$.

Portanto o estimador de quadrados mínimos $\hat{\boldsymbol{\beta}}^{\mathrm{o}\,*}$ satisfaz $\hat{\boldsymbol{\beta}}^{\mathrm{o}\,*} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^{\mathrm{r}}(\lambda_2)$, isto é, é exatamente o valor obtido pela reparametrização da estimativa Ridge da regressão original $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

A estimativa LASSO para a regressão $\mathbf{y}^* = \mathrm{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ é obtida minimizando a Lagrangeana

$$\mathrm{L}\,(\gamma,\boldsymbol{\beta}^*) = \|\mathbf{y} - \mathrm{X}^*\boldsymbol{\beta}^*\|^2 + \gamma\|\boldsymbol{\beta}^*\|_1,$$

isto é, $\boldsymbol{\beta}^{\mathrm{lasso}\,*} = \underset{\boldsymbol{\beta}^*}{\arg\min}\,\mathrm{L}\,(\gamma,\boldsymbol{\beta}^*)$, em que $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$.

Como

$$\begin{aligned}
\|\mathbf{y}^* - \mathrm{X}^*\boldsymbol{\beta}^*\|^2 + \gamma\|\boldsymbol{\beta}^*\|_1 &= \left\| \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} - \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathrm{X} \\ \sqrt{\lambda_2}\,\mathrm{I} \end{pmatrix} \sqrt{1+\lambda_2}\boldsymbol{\beta} \right\|^2 \\
&\quad + \gamma \left\| \sqrt{1+\lambda_2}\boldsymbol{\beta} \right\|_1 \\
&= \left\| \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} - \frac{1}{1} \begin{pmatrix} \mathrm{X}\boldsymbol{\beta} \\ \sqrt{\lambda_2}\boldsymbol{\beta} \end{pmatrix} \right\|^2 + \sqrt{1+\lambda_2}\gamma\|\boldsymbol{\beta}\|_1 \\
&= \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2 + \sqrt{1+\lambda_2}\gamma\|\boldsymbol{\beta}\|_1 \\
&= \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1,
\end{aligned}$$

tem-se que o estimador LASSO $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}\,*}$ é exatamente o estimador $\hat{\boldsymbol{\beta}}_{eni}$ pois são definidos pela minimização da mesma função. Em termos dos parâmetros originais, $\hat{\boldsymbol{\beta}}_{eni} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^{\mathrm{lasso}\,*}$. Desta forma, o cálculo do estimador *elastic net* ingênuo pode ser obtido como o estimador LASSO de um sistema aumentado.

Para exemplificar, vamos calcular a estimativa *elastic net* ingênua para o caso ortogonal $\mathrm{X}'\mathrm{X} = \mathrm{I}$. Conforme demonstrado (fórmula 4.1), o estimador LASSO para uma regressão ortogonal é dado por

$$\left(\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}\right)_{\mathrm{j}} = \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_{\mathrm{j}} \left(\left|\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_{\mathrm{j}}\right| - \gamma\right)^{+}.$$

A regressão com os dados aumentados satisfaz

$$\begin{aligned}
\mathrm{X}^{*} &= \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathrm{X}\\ \sqrt{\lambda_2}\,\mathrm{I}\end{pmatrix} \Rightarrow \\
\left(\mathrm{X}^{*}\right)'\left(\mathrm{X}^{*}\right) &= \frac{1}{1+\lambda_2}\left(\mathrm{X}', \sqrt{\lambda_2}\mathrm{I}\right)\begin{pmatrix}\mathrm{X}\\ \sqrt{\lambda_2}\mathrm{I}\end{pmatrix} \\
&= \frac{1}{1+\lambda_2}\left(\mathrm{X}'\mathrm{X}+\lambda_2\mathrm{I}\right) \\
&= \frac{1}{1+\lambda_2}\left(\mathrm{I}+\lambda_2\mathrm{I}\right) \\
&= \mathrm{I}.
\end{aligned}$$

Portanto o sistema aumentado também é ortogonal. O LASSO do sistema aumentado é

$$\left(\hat{\boldsymbol{\beta}}^{\text{lasso }*}\right)_j = \left(\left|\left(\hat{\boldsymbol{\beta}}^{\text{ols }*}\right)_j\right| - \gamma\right)^{+} \text{sinal}\left(\hat{\boldsymbol{\beta}}^{\text{ols }*}\right)_{\text{j}}.$$

Como demonstrado,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{ols }*} &= \sqrt{1+\lambda_2}\,\hat{\boldsymbol{\beta}}_R(\lambda_2) \\
&= \sqrt{1+\lambda_2}\left(\mathrm{X}'\mathrm{X}+\lambda_2\mathrm{I}\right)^{-1}\mathrm{X}'\mathbf{y} \\
&= \sqrt{1+\lambda_2}((1+\lambda_2)\mathrm{I})^{-1}\mathrm{X}'\mathbf{y} \\
&= \frac{1}{\sqrt{1+\lambda_2}}\mathrm{X}'\mathbf{y}.
\end{aligned}$$

Mas $\mathrm{X}'\mathbf{y}$ é o estimador de quadrados mínimos da regressão original e portanto $\hat{\boldsymbol{\beta}}^{\text{ols }*} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^{\text{ols}}$. Como o estimador *elastic net* ingênuo satisfaz $\hat{\boldsymbol{\beta}}_{eni} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^{\text{lasso }*}$, então

$$\begin{aligned}
\left(\hat{\boldsymbol{\beta}}_{\text{eni}}\right)_j &= \frac{1}{\sqrt{1+\lambda_2}}\left(\hat{\boldsymbol{\beta}}^{\text{lasso }*}\right)_j \\
&= \frac{1}{\sqrt{1+\lambda_2}}\left(\left|\left(\hat{\boldsymbol{\beta}}^{\text{ols }*}\right)_j\right| - \gamma\right)^{+} \text{sinal}\left(\hat{\boldsymbol{\beta}}^{\text{ols }*}\right)_{\text{j}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{1+\lambda_2}}\left(\left|\frac{1}{\sqrt{1+\lambda_2}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_j\right| - \gamma\right)^+ \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_{\mathrm{j}} \\
&= \frac{\left(\left|\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_j\right| - \sqrt{(1+\lambda_2)}\gamma\right)^+}{1+\lambda_2} \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_{\mathrm{j}} \\
&= \frac{\left(\left|\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_j\right| - \frac{\lambda_1}{2}\right)^+}{1+\lambda_2} \mathrm{sinal}\left(\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right)_{\mathrm{j}},
\end{aligned}
$$

Zou (2005, p. 305).

A grande vantagem do método *elasticnet* em relação ao método LASSO é que se tem a propriedade de efeito de grupo, isto é, variáveis altamente correlacionadas têm alta probabilidade de gerar estimativas relacionadas a estas covariáveis com valores próximos. Esta propriedade decorre do fato que a penalidade empregada no método *elasticnet* define um modelo $\mathrm{K_p}$ estritamente convexo, ao contrário do método LASSO que define um modelo $\mathrm{K_p}$ apenas convexo.

A diferença entre uma penalização convexa e uma estritamente convexa é explicitada na Proposição 5. Seja $\mathrm{J}(\,\cdot\,)$ uma função positiva e o problema de minimização

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda \mathrm{J}\left(\boldsymbol{\beta}\right) \right\}.$$

Tem-se

Proposição 5. *Se* $\mathbf{x}_{\mathrm{i}} = \mathbf{x}_{\mathrm{j}}$, $\mathrm{i}, \mathrm{j} \in \{1, \ldots, \mathrm{p}\}$,

1. *Se* $\mathrm{J}(\cdot)$ *é estritamente convexa, então* $\hat{\beta}_{\mathrm{i}} = \hat{\beta}_{\mathrm{j}}$, $\forall \lambda > 0$.

2. *Se* $\mathrm{J}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$, *portanto convexa mas não estritamente convexa, então* $\hat{\beta}_{\mathrm{i}}\hat{\beta}_{\mathrm{j}} \geq 0$ *e*

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & k \neq i, j \\ \left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right)\mathrm{s}, & \mathrm{k} = \mathrm{i} \\ \left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right)(1-\mathrm{s}), & k = j \end{cases}$$

é uma outra solução do problema de minimização para $\mathrm{s} \in [0,1]$.

Demonstração. 1) Se J é estritamente convexa, então $\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda \mathrm{J}(\boldsymbol{\beta})$ também o é e, portanto, admite um único ponto de mínimo $\hat{\boldsymbol{\beta}}$. Seja agora o caso em que $\mathbf{x_i} = \mathbf{x_j}$. Definindo um novo vetor $\hat{\boldsymbol{\beta}}^*$ da forma

$$\hat{\beta}^*_{\mathrm{k}} = \begin{cases} \hat{\beta}_{\mathrm{k}} & \text{se } \mathrm{k} \neq \mathrm{i}, \mathrm{j} \\ \frac{1}{2}\left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right) & \text{se } \mathrm{k} = \mathrm{i}, \mathrm{j} \end{cases}.$$

Como as i-ésima e j-ésima colunas de X são iguais, tem-se que $\mathrm{X}\hat{\boldsymbol{\beta}}^* = \mathrm{X}\hat{\boldsymbol{\beta}}$, de onde segue que $\hat{\boldsymbol{\beta}}^*$ também é um mínimo e pela unicidade $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}$, o que implica que $\hat{\beta}_{\mathrm{i}} = \hat{\beta}_{\mathrm{j}}$.

2) Suponha $\hat{\beta}_{\mathrm{i}}\,\hat{\beta}_{\mathrm{j}} < 0$ e considere novamente o vetor $\hat{\boldsymbol{\beta}}^*$. Assim,

$$\left(\hat{\beta}^*_{\mathrm{j}}\right)^2 = \left(\hat{\beta}^*_{\mathrm{i}}\right)^2 = \hat{\beta}^*_{\mathrm{i}}\,\hat{\beta}^*_{\mathrm{j}} = \frac{1}{2}\left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right)\frac{1}{2}\left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right)$$
$$= \frac{1}{4}\left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right)^2 = \frac{1}{4}\left(\hat{\beta}^2_{\mathrm{i}} + 2\hat{\beta}_{\mathrm{i}}\,\hat{\beta}_{\mathrm{j}} + \hat{\beta}^2_{\mathrm{j}}\right) \leq \frac{1}{4}\left(\hat{\beta}^2_{\mathrm{i}} + \hat{\beta}^2_{\mathrm{j}}\right).$$

Portanto,

$$\left\|\hat{\boldsymbol{\beta}}^*\right\|_1 = \sum_{\mathrm{s}=1}^{\mathrm{p}} \left|\hat{\beta}^*_{\mathrm{s}}\right| = \sum_{\mathrm{s}\neq \mathrm{i},\mathrm{j}}^{\mathrm{p}} \left|\hat{\beta}^*_{\mathrm{s}}\right| + \left|\hat{\beta}^*_{\mathrm{i}}\right| + \left|\hat{\beta}^*_{\mathrm{j}}\right| = \sum_{\mathrm{s}\neq \mathrm{i},\mathrm{j}}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{s}}\right| + 2\left|\hat{\beta}^*_{\mathrm{i}}\right|.$$

Como $\hat{\beta}^*_{\mathrm{i}} = \frac{1}{2}\left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right)$, se $\hat{\beta}_{\mathrm{i}}\,\hat{\beta}_{\mathrm{j}} < 0$, $\hat{\beta}_{\mathrm{i}}$ e $\hat{\beta}_{\mathrm{i}}$ possuem sinais opostos e $\left|\hat{\beta}^*_{\mathrm{i}}\right| = \frac{1}{2}\left|\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right| < \frac{1}{2}\left(\left|\hat{\beta}_{\mathrm{i}}\right| + \left|\hat{\beta}_{\mathrm{j}}\right|\right)$. Logo,

$$\left\|\hat{\boldsymbol{\beta}}^*\right\|_1 < \sum_{\mathrm{s}\neq \mathrm{i},\mathrm{j}}^{\mathrm{p}} \left|\hat{\beta}_{\mathrm{s}}\right| + 2\left(\frac{1}{2}\left(\left|\hat{\beta}_{\mathrm{i}}\right| + \left|\hat{\beta}_{\mathrm{j}}\right|\right)\right) = \left\|\hat{\boldsymbol{\beta}}\right\|_1.$$

Tal fato não pode ocorrer pois $\hat{\boldsymbol{\beta}}^*$ e $\hat{\boldsymbol{\beta}}$ minimizam $\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{\beta}\|_1$ e portanto $\hat{\beta}_{\mathrm{i}}\,\hat{\beta}_{\mathrm{j}} \geq 0$.

Observando que se define

$$\hat{\beta}^*_{\mathrm{k}} = \begin{cases} \hat{\beta}_{\mathrm{k}} & , \ \mathrm{k} \neq \mathrm{i}, \mathrm{j} \\ \left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right) \mathrm{s} & , \ \mathrm{k} = \mathrm{i} \\ \left(\hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}}\right) (1 - \mathrm{s}) & , \ \mathrm{k} = \mathrm{j} \end{cases} , \ \mathrm{s} \in [0, 1] ,$$

então

$$\left\| \hat{\boldsymbol{\beta}}^* \right\|_1 = \sum_{\mathrm{s} \neq \mathrm{i},\mathrm{j}}^{\mathrm{p}} \left| \hat{\beta}_{\mathrm{s}} \right| + \left| \hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}} \right| \mathrm{s} + \left| \hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}} \right| (1 - \mathrm{s})$$

$$= \sum_{\mathrm{s} \neq \mathrm{i},\mathrm{j}}^{\mathrm{p}} \left| \hat{\beta}_{\mathrm{s}} \right| + \left| \hat{\beta}_{\mathrm{i}} + \hat{\beta}_{\mathrm{j}} \right| = \sum_{\mathrm{s} \neq \mathrm{i},\mathrm{j}}^{\mathrm{p}} \left| \hat{\beta}_{\mathrm{s}} \right| + \left| \hat{\beta}_{\mathrm{i}} \right| + \left| \hat{\beta}_{\mathrm{j}} \right| = \left\| \hat{\boldsymbol{\beta}} \right\|_1 ,$$

pois $\hat{\beta}_{\mathrm{i}}$ e $\hat{\beta}_{\mathrm{j}}$ possuem o mesmo sinal. Portanto $\hat{\boldsymbol{\beta}}^*_{\mathrm{s}}$ também é solução do problema de minimização. □

Como o caso $\mathbf{x}_{\mathrm{i}} = \mathbf{x}_{\mathrm{j}}$ não é de interesse prático, um resultado de maior aplicabilidade vai garantir uma cota para a diferença entre as estimativas $\left(\hat{\boldsymbol{\beta}}_{eni}\right)_{\mathrm{i}}$ e $\left(\hat{\boldsymbol{\beta}}_{eni}\right)_{\mathrm{j}}$ em termos da correlação amostral $\rho = \mathbf{x}'_{\mathrm{i}}\mathbf{x}_{\mathrm{j}}$.

Teorema 1. *Se $\hat{\boldsymbol{\beta}}_{\mathrm{eni}}(\lambda_1, \lambda_2)$ é o vetor de estimativas elasticnet ingênua e se*

$$\left(\hat{\boldsymbol{\beta}}_{\mathrm{eni}}\right)_{\mathrm{i}}(\lambda_1, \lambda_2)\left(\hat{\boldsymbol{\beta}}_{\mathrm{eni}}\right)_{\mathrm{j}}(\lambda_1, \lambda_2) > 0,$$

então

$$\frac{1}{\|\mathbf{y}\|_1} \left| \left(\hat{\boldsymbol{\beta}}_{\mathrm{eni}}\right)_{\mathrm{i}}(\lambda_1, \lambda_2) - \left(\hat{\boldsymbol{\beta}}_{\mathrm{eni}}\right)_{\mathrm{j}}(\lambda_1, \lambda_2) \right| \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho)}.$$

Demonstração. Observando que a diferenciabilidade da função $\mathrm{L}(\lambda_1, \lambda_2, \boldsymbol{\beta})$ é garantida apenas fora das intersecções com os eixos coordenados, então

$$\frac{\partial}{\partial \boldsymbol{\beta}_{\mathrm{k}}} \mathrm{L}(\lambda_1, \lambda_2, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)} = 0$$

faz sentido se $\hat{\boldsymbol{\beta}}_{\mathrm{k}}(\lambda_1, \lambda_2) \neq 0$. Com essa suposição, a derivada da Lagrangeana

em relação a $\boldsymbol{\beta}_{\mathrm{i}}$ é

$$-2\mathrm{X}'\left\{\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\beta}}\left(\lambda_1,\lambda_2\right)\right\}+\lambda_1 \operatorname{sinal}\hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)+2\lambda_2\hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)=0.$$

Subtraindo esta equação para i e j, se tem

$$\left(\mathbf{x}_{\mathrm{j}}-\mathbf{x}_{\mathrm{i}}\right)'\left\{\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\beta}}\left(\lambda_1,\lambda_2\right)\right\}+\lambda_2\left(\hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)-\hat{\boldsymbol{\beta}}_{\mathrm{j}}\left(\lambda_1,\lambda_2\right)\right)=0$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)-\hat{\boldsymbol{\beta}}_{\mathrm{j}}\left(\lambda_1,\lambda_2\right)=\frac{1}{\lambda_2}\left(\mathbf{x}'_{\mathrm{i}}-\mathbf{x}'_{\mathrm{j}}\right)\left(\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\beta}}\right). \tag{5.1}$$

Como

$$\begin{aligned}\left\|\mathbf{x}_{\mathrm{i}}-\mathbf{x}_{\mathrm{j}}\right\|^2 &= \left\langle\mathbf{x}_{\mathrm{i}}-\mathbf{x}_{\mathrm{j}},\mathbf{x}_{\mathrm{i}}-\mathbf{x}_{\mathrm{j}}\right\rangle\\ &= \left\|\mathbf{x}_{\mathrm{i}}\right\|^2+\left\|\mathbf{x}_{\mathrm{j}}\right\|^2-2\mathbf{x}'_{\mathrm{i}}\,\mathbf{x}_{\mathrm{j}}\\ &= 1+1-2\rho\\ &= 2\left(1-\rho\right).\end{aligned}$$

e também

$$\mathrm{L}\left(\lambda_1,\lambda_2,\hat{\boldsymbol{\beta}}\right)\leq \mathrm{L}\left(\lambda_1,\lambda_2,\boldsymbol{\beta}=0\right)=\left\|\mathbf{y}\right\|^2$$

$$\Rightarrow \left\|\mathbf{y}-\mathrm{X}\hat{\boldsymbol{\beta}}\right\|^2+\lambda_1\left\|\hat{\boldsymbol{\beta}}\right\|_1+\lambda_2\left\|\hat{\boldsymbol{\beta}}\right\|^2\leq\left\|\mathbf{y}\right\|^2,$$

substituindo na equação 5.1 e utilizando-se do fato de que $\hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)$ e $\hat{\boldsymbol{\beta}}_{\mathrm{j}}\left(\lambda_1,\lambda_2\right)$ têm o mesmo sinal, tem-se $\left|\hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)-\hat{\boldsymbol{\beta}}_{\mathrm{j}}\left(\lambda_1,\lambda_2\right)\right|\leq\frac{1}{\lambda_2}2\left(1-\rho\right)\left\|\mathbf{y}\right\|^2$, e portanto

$$\mathrm{D}_{\lambda_1,\lambda_2}\left(i,j\right)=\frac{\left|\hat{\boldsymbol{\beta}}_{\mathrm{i}}\left(\lambda_1,\lambda_2\right)-\hat{\boldsymbol{\beta}}_{\mathrm{j}}\left(\lambda_1,\lambda_2\right)\right|}{\left\|\mathbf{y}\right\|_1}\leq\frac{1}{\lambda_2}2\left(1-\rho\right)\frac{\left\|\mathbf{y}\right\|^2}{\left\|\mathbf{y}\right\|_1}\leq\frac{2\left(1-\rho\right)}{\lambda_2}$$

□

Segue do Teorema 1 que se $\mathbf{x}_{\mathrm{i}}$ e $\mathbf{x}_{\mathrm{j}}$ são altamente correlacionados, então a diferença das estimativas dos parâmetros relativos a estas covariáveis é próxima de

zero. Esta propriedade é que dá ao estimador sua propriedade de seleção por grupo, isto é, grupos de variáveis altamente correlacionadas tendem a ter estimativas correspondentes próximas. Esta propriedade é muito interessante em aplicações. O LASSO, em geral, não possui tal propriedade.

Como observado anteriormente, o estimador *elastic net* ingênuo é obtido por um procedimento em dois estágios: uma estimativa RIDGE e em seguida um encolhimento tipo LASSO. Portanto, se tem um duplo encolhimento e estudos de simulação comprovam que tal fato afeta a performance do estimador. Para que se possa corrigir tal situação, o estimador *elastic net* é definido como um reescalonamento do estimador *elastic net* ingênuo $\hat{\boldsymbol{\beta}}_{\text{en}} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{\text{eni}}$. O teorema seguinte fornece a forma explícita para o estimador.

Teorema 2. *O estimador elastic net satisfaz*

$$\hat{\boldsymbol{\beta}}_{\text{en}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \boldsymbol{\beta}' \frac{(\mathrm{X}'\mathrm{X}+\lambda_2\mathrm{I})}{1+\lambda_2}\boldsymbol{\beta} - 2\mathbf{y}'\mathrm{X}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1.$$

Observe que a estimativa LASSO é dada por

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{lasso}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \|\mathbf{y}-\mathrm{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \left\{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathrm{X}\boldsymbol{\beta} - (\mathrm{X}\boldsymbol{\beta})'\mathbf{y} + (\mathrm{X}\boldsymbol{\beta})'(\mathrm{X}\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1\right\} \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \left\{\mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta} - 2\mathbf{y}'\mathrm{X}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1\right\} \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \left\{\boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta} - 2\mathbf{y}'\mathrm{X}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1\right\}.
\end{aligned}$$

Desta forma fica claro que o Teorema 2 nos garante que o estimador *elastic net* é uma versão estabilizada do estimador LASSO.

Demonstração. $(\mathbf{y}^*, \mathrm{X}^*)$ é o sistema aumentado descrito anteriormente e

$$\hat{\boldsymbol{\beta}}_{\text{en}} = \sqrt{(1+\lambda_2)}\hat{\boldsymbol{\beta}}^* = \sqrt{(1+\lambda_2)}\,\underset{\boldsymbol{\beta}^*}{\operatorname{argmin}} \|\mathbf{y}^* - \mathrm{X}\boldsymbol{\beta}^*\|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}}\|\boldsymbol{\beta}^*\|_1,$$

em que $\boldsymbol{\beta}^*$ é uma reparametrização dada por $\boldsymbol{\beta}^* = \sqrt{(1+\lambda_2)}\boldsymbol{\beta}$. Portanto,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{en}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{y} - \mathrm{X}^* \frac{\boldsymbol{\beta}}{\sqrt{1+\lambda_2}} \right\|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \left\| \frac{\boldsymbol{\beta}}{\sqrt{1+\lambda_2}} \right\|_1 \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathrm{X}^* \frac{\boldsymbol{\beta}}{\sqrt{1+\lambda_2}} + \boldsymbol{\beta}'(\mathrm{X}^*)'\mathrm{X}^*\boldsymbol{\beta} + \frac{\lambda_1}{1+\lambda_2}\|\boldsymbol{\beta}\|_1 \right) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\boldsymbol{\beta}'(\mathrm{X}^*)'\mathrm{X}^*\boldsymbol{\beta} - 2\mathbf{y}'\mathrm{X}^* \frac{\boldsymbol{\beta}}{\sqrt{1+\lambda_2}} + \frac{\lambda_1}{1+\lambda_2}\|\boldsymbol{\beta}\|_1 \right) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\boldsymbol{\beta}' \frac{(\mathrm{X}'\mathrm{X}+\lambda_2\mathrm{I})}{(1+\lambda_2)^2}\boldsymbol{\beta} - \frac{2\mathbf{y}'\mathrm{X}\boldsymbol{\beta}}{1+\lambda_2} + \frac{\lambda_1}{1+\lambda_2}\|\boldsymbol{\beta}\|_1 \right) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\boldsymbol{\beta}' \frac{(\mathrm{X}'\mathrm{X}+\lambda_2\mathrm{I})}{1+\lambda_2}\boldsymbol{\beta} - 2\mathbf{y}'\mathrm{X}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1 \right).
\end{aligned}$$

□

6 Regressão de Ângulos Mínimos

Métodos de seleção de modelos, isto é, seleção de covariáveis, são clássicos na teoria dos modelos lineares. Pode-se citar entre eles: *Forward selection*, *Backward selection*, *All subset selection*, e mais recentemente os métodos LASSO e *Forward stagewise linear regression*. Este último deu origem a um método computacionalmente muito mais simples denominado *Least Angle Regression*. Estes métodos podem ser vistos como um aprimoramento de um método mais elementar denominado *Forward stepwise regression*.

6.1 Forward Stagewise Regression

O método *Forward stepwise* consiste em primeiramente se escolher dentre as covariáveis, isto é, entre os vetores $\mathbf{x}_i$ (colunas da matriz X) aquele que tem a maior correlação com o vetor de dados $\mathbf{y}$. A Figura 41 representa o vetor $\mathbf{y}$ e os vetores das covariáveis $\mathbf{x}_1, \ldots, \mathbf{x}_p$.

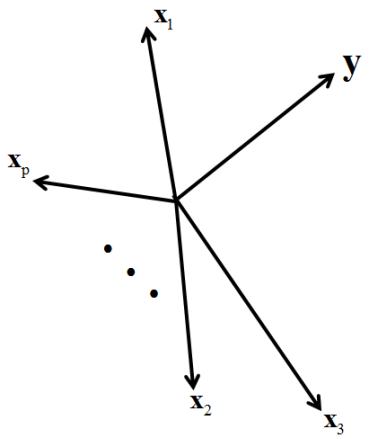


Figura 41 Vetor de dados e das covariáveis.

De forma a simplificar, é possível considerar a matriz X normalizada, isto é, $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}^2 = 1$ e também $\sum_{i=1}^{n} y_i = 0$. Com isso, os vetores das covariáveis estão representados em uma hiperesfera de raio unitário (ver Figura 42).

A correlação amostral é dada por $\mathbf{x'}_i\mathbf{y} = \|\mathbf{x}_i\| \, \|\mathbf{y}\| \cos\theta = \|\mathbf{y}\| \cos\theta = P_{\mathbf{x}_i}\mathbf{y}$ (ver representação na Figura 43).

Pode-se supor que o vetor $\mathbf{y} - P_{\mathbf{x}_i}\mathbf{y}$ é a parte dos dados $\mathbf{y}$ não explicada pela covariável $\mathbf{x}_i$. A ideia então é explicar o vetor $\mathbf{y} - P_{\mathbf{x}_i}\mathbf{y}$ em termos das outras covariáveis $\mathbf{x}_j$, $j \neq i$. Para tanto basta considerar as componentes dos vetores $\mathbf{x}_j$ que são perpendiculares à $\mathbf{x}_i$ (ver Figura 44).

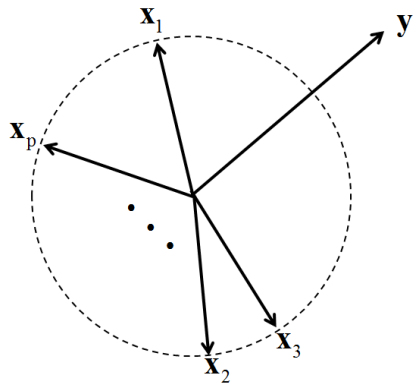


Figura 42 Vetores normalizados das covariáveis.

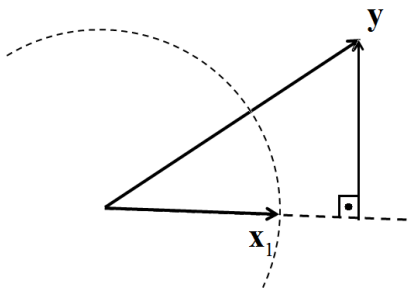


Figura 43 Geometria da correlação amostral.

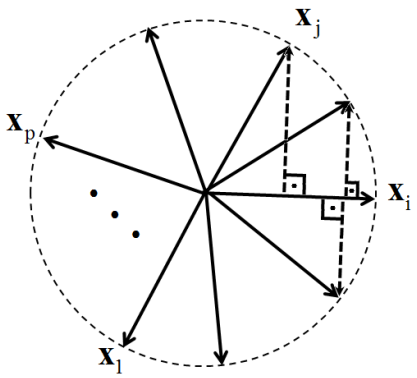


Figura 44 Normalização das covariáveis.

O processo agora é repetido considerando como dados o ortogonal do vetor $\mathbf{y}$ em relação ao vetor $\mathbf{x}_{\mathrm{i}}$ e como vetores das covariáveis os ortogonais de $\mathbf{x}_{\mathrm{j}}$ em relação à $\mathbf{x}_{\mathrm{i}}$. Novamente se escolhe a covariável que apresenta maior correlação amostral com $\mathbf{y} - \mathrm{P}_{\mathbf{x}_{\mathrm{i}}}\mathbf{y}$. Desta forma se tem uma sequência encaixada de modelos. O vetor de regressão é obtido da forma: Supondo por simplicidade que o modelo tenha escolhido na sequência as variáveis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{k}}$, se teria então

$$\begin{aligned} \mathbf{y} &= \mathrm{P}_{\mathbf{x}_1}\mathbf{y} + (\mathbf{y} - \mathrm{P}_{\mathbf{x}_1}\mathbf{y}) \\ &= \mathrm{P}_{\mathbf{x}_1}\mathbf{y} + \mathrm{P}_{\mathbf{x}_2}(\mathbf{y} - \mathrm{P}_{\mathbf{x}_1}\mathbf{y}) + (\mathbf{y} - \mathrm{P}_{\mathbf{x}_1}\mathbf{y} - (\mathrm{P}_{\mathbf{x}_2}(\mathbf{y} - \mathrm{P}_{\mathbf{x}_1}\mathbf{y}))). \end{aligned}$$

O processo é repetido e o vetor de regressão é obtido desprezando-se o último resíduo.

Duas covariáveis altamente correlacionadas devem fazer parte do modelo ou pelo princípio da parcimônia é razoável que apenas uma delas esteja presente no modelo? Posteriormente, por algum critério, um destes modelos é então selecionado. Este processo tende a ser agressivo (EFRON et al., 2004) no sentido de que logo na segunda iteração uma covariável $\mathbf{x}_{\mathrm{j}}$, altamente correlacionada com $\mathbf{x}_{\mathrm{i}}$, não seja escolhida e fique excluída do modelo obtido. Geometricamente tal fato se justifica pois supondo $\mathbf{x}_{\mathrm{j}}$ altamente correlacionada com $\mathbf{x}_{\mathrm{i}}$ então os dois vetores estão próximos. Como $\mathbf{x}_{\mathrm{i}}$ e $\mathbf{x}_{\mathrm{j}}$ estão próximos, então $\mathbf{y} - \mathrm{P}_{\mathbf{x}_{\mathrm{i}}}\mathbf{y}$ e $\mathbf{y} - \mathrm{P}_{\mathbf{x}_{\mathrm{j}}}\mathbf{y}$ também estarão próximos, mas $\mathbf{x}_{\mathrm{j}} - \mathrm{P}_{\mathbf{x}_{\mathrm{i}}}\mathbf{x}_{\mathrm{j}}$ é um vetor tangente à hiperesfera no ponto $\mathbf{x}_{\mathrm{i}}$, e portanto não necessariamente muito correlacionado com $\mathbf{y} - \mathrm{P}_{\mathbf{x}_{\mathrm{i}}}\mathbf{y}$ (Figura 45).

Em razão desta deficiência, foi proposto um outro algoritmo de seleção, mais cauteloso, denominado *Forward Stagewise*. A descrição do algoritmo é como se segue. Seja $\hat{\boldsymbol{\mu}}$ a estimativa de $\mathrm{E}[\mathbf{y}]$ em um passo do algoritmo. No

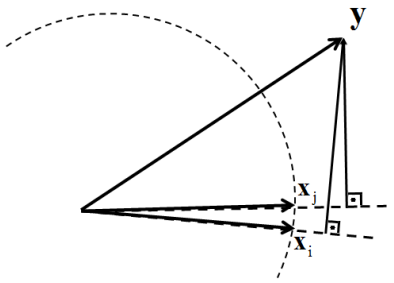


Figura 45 Stepwise e variáveis altamente correlacionadas.

primeiro passo se toma $\hat{\boldsymbol{\mu}} = 0$. Considere o vetor de correlações amostrais (de fato proporcional às correlações) $\hat{\mathbf{c}} = \mathrm{c}\,(\hat{\boldsymbol{\mu}}) = (\hat{\mathrm{c}}_1, \ldots, \hat{\mathrm{c}}_\mathrm{p}) = \mathrm{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}) = (\mathbf{x}'_1(\mathbf{y} - \hat{\boldsymbol{\mu}}), \ldots, \mathbf{x}'_\mathrm{p}(\mathbf{y} - \hat{\boldsymbol{\mu}}))$ e seja $\hat{\mathrm{j}} = \arg\max |\hat{\mathrm{c}}_\mathrm{j}|$. O estimador $\hat{\boldsymbol{\mu}}$ é então atualizado por

$$\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + \varepsilon \mathrm{sinal}(\hat{\mathrm{c}}_\mathrm{j})\mathbf{x}_{\hat{\mathrm{j}}},$$

em que ε é um parâmetro de controle do processo. Observe que não se está fazendo distinção entre positivamente e negativamente correlacionado. Do ponto de vista das aplicações, se quer selecionar covariáveis que estejam altamente correlacionadas, positiva ou negativamente com a variável resposta.

O índice $\hat{\mathrm{j}}$ escolhido é aquele determinado pelo menor ou maior ângulo entre as covariáveis $\mathbf{x}_\mathrm{i}$ relativo à $\mathbf{y} - \hat{\boldsymbol{\mu}}$.

No caso em que se toma $\varepsilon = |\hat{\mathrm{c}}_\mathrm{j}|$ ocorre que o processo se reduz ao *stepwise*. De fato, no primeiro passo do algoritmo $\hat{\boldsymbol{\mu}} = 0$ e

$$\hat{\boldsymbol{\mu}} \leftarrow 0 + \varepsilon \mathrm{sinal}(\hat{\mathrm{c}}_\mathrm{j})\mathbf{x}_\mathrm{j} = |\hat{\mathrm{c}}_\mathrm{j}|\,\mathrm{sinal}(\hat{\mathrm{c}}_\mathrm{j})\mathbf{x}_\mathrm{j} = \hat{\mathrm{c}}_\mathrm{j}\mathbf{x}_\mathrm{j} = \mathbf{x}'_\mathrm{j}(\mathbf{y} - 0)\mathbf{x}_\mathrm{j} = \mathrm{P}_{\mathbf{x}_\mathrm{j}}\mathbf{y},$$

que é justamente a projeção de $\mathbf{y}$ no vetor $\mathbf{x}_\mathrm{j}$ a menos do sinal. No passo seguinte, $\hat{\mathrm{c}} = \mathbf{x}'(\mathbf{y} - \mathrm{P}_{\mathbf{x}_\mathrm{i}}\mathbf{y})$ e portanto está se calculando a correlação das covariáveis com um vetor ortogonal à $\mathbf{x}_\mathrm{i}$, e o processo se reduz ao *Forward Stepwise*. Portanto, o

valor de ε deve ser pequeno para que o processo seja capaz de detectar covariáveis muito correlacionadas. Assim, antes de se escolher uma nova covariável, vários passos ocorrem para a mesma covariável. Ver Figura 46.

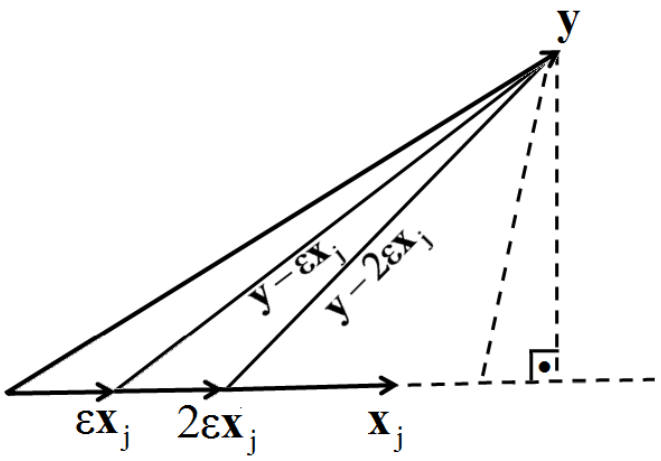


Figura 46 Algoritmo *stagewise*.

6.2 O Algoritmo LARS

O algoritmo LARS segue o mesmo procedimento do método *Forward Stagewise*, só que com uma ideia matemática simples, que consiste na escolha adequada para o valor de ε que define o passo do algoritmo. O número de passos é drasticamente reduzido igualando-se ao número de covariáveis. Primeiramente vamos descrever o algoritmo para o caso simples de 2 covariáveis $\mathrm{X} = (\mathbf{x}_1,\mathbf{x}_2)$.

O vetor de correlações amostrais é

$$\mathbf{c}(\hat{\boldsymbol{\mu}}) = (\mathrm{c}_1(\hat{\boldsymbol{\mu}}), \mathrm{c}_2(\hat{\boldsymbol{\mu}})) = \mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \left(\mathbf{x}'_1(\mathbf{y} - \hat{\boldsymbol{\mu}}), \mathbf{x}'_2(\mathbf{y} - \hat{\boldsymbol{\mu}})\right).$$

Inicialmente $\boldsymbol{\mu}_0 = 0$. Para fixar a construção suponha que o ângulo entre $\mathbf{y}$ e $\mathbf{x}_1$ seja menor do que com $\mathbf{x}_2$ (Figura 47), isto é, $\mathrm{c}_1(\hat{\boldsymbol{\mu}}_0) > \mathrm{c}_2(\hat{\boldsymbol{\mu}}_0)$. Logo o algoritmo aumenta $\boldsymbol{\mu}_0$ na direção de $\mathbf{x}_1$ para $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\gamma}_1\mathbf{x}_1$. A ideia então é escolher

o valor de $\hat{\gamma}_1$ de forma adequada, tal que $\mathbf{y} - \hat{\boldsymbol{\mu}}_1$ seja igualmente correlacionado com $\mathbf{x}_1$ e $\mathbf{x}_2$. Isto ocorre se $\mathbf{y} - \hat{\boldsymbol{\mu}}_1$ é equiangular com $\mathbf{x}_1$ e $\mathbf{x}_2$. Observe que pela ortogonalidade é possível substituir $\mathbf{y}$ por $\bar{\mathbf{y}}_2 = \mathrm{P}_{\mathrm{Im(X)}}\mathbf{y}$. Portanto se quer que $\bar{\mathbf{y}}_2 - \hat{\boldsymbol{\mu}}_1$ seja equiangular com $\mathbf{x}_1$ e $\mathbf{x}_2$, isto é,

$$\mathbf{x'}_1(\bar{\mathbf{y}}_2 - \hat{\gamma}_1\mathbf{x}_1) = \mathbf{x'}_2(\bar{\mathbf{y}}_2 - \hat{\gamma}_1\mathbf{x}_1) \Rightarrow \hat{\gamma}_1 = \frac{(\mathbf{x'}_2 - \mathbf{x'}_1)\mathbf{y}}{\mathbf{x'}_2\mathbf{x}_1 - \mathbf{x'}_1\mathbf{x}_1}.$$

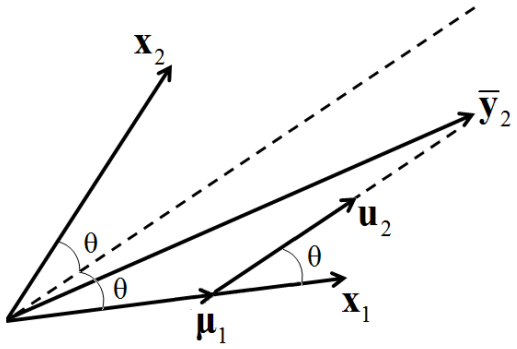


Figura 47 Geometria do LARS para a dimensão 2.

Se $\mathbf{u_2}$ é o vetor unitário na direção equiangular, o próximo passo do algoritmo atualiza $\boldsymbol{\mu}$ da forma $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\gamma}_2\mathbf{u_2}$. No caso bidimensional, $\hat{\gamma}_2$ é escolhido tal que $\hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{y}}_2$.

Se a dimensão fosse maior do que 2, isto é, com uma nova covariável $\mathbf{x}_3$ e vetor equiangular $\mathbf{u}_3$, $\hat{\gamma}_2$ seria menor e se teria uma outra alteração de direção. O valor de $\hat{\gamma}_2$ é obtido pelas igualdades

$$\mathbf{x'}_1\left(\bar{\mathbf{y}}_3 - \hat{\gamma}_2\hat{\boldsymbol{\mu}}_2\right) = \mathbf{x'}_2\left(\bar{\mathbf{y}}_3 - \hat{\gamma}_2\hat{\boldsymbol{\mu}}_2\right) = \mathbf{x'}_3\left(\bar{\mathbf{y}}_3 - \hat{\gamma}_2\hat{\boldsymbol{\mu}}_2\right),$$

e o valor de $\hat{\gamma}_3$ é dado por $\hat{\boldsymbol{\mu}}_3 = \bar{\mathbf{y}}_3 = \hat{\boldsymbol{\mu}}_2 + \hat{\gamma}_3\mathbf{u}_3$, conforme descrito na Figura 48.

O procedimento algébrico para a obtenção de um vetor $\mathbf{u}$ em $\mathbb{R}^{\mathrm{n}}$ equiangular com os vetores colunas $\mathbf{x}_{\mathrm{i}}$ da matriz X será descrito com detalhes. Ob-

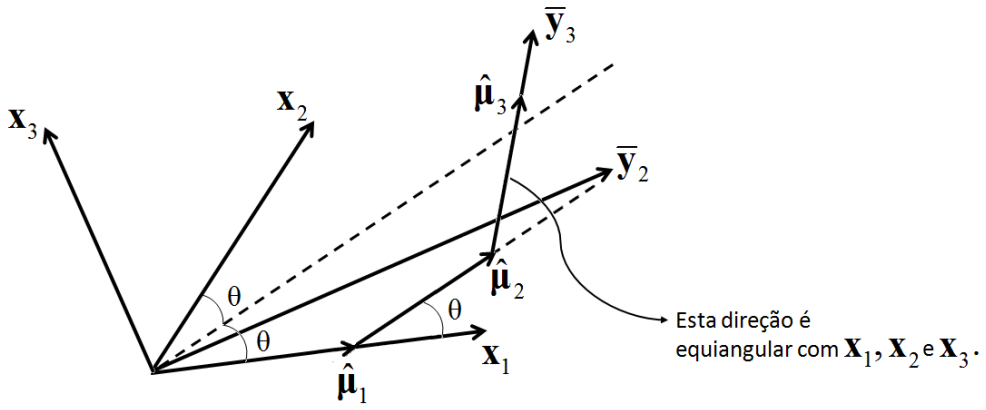


Figura 48 Geometria do LARS para a dimensão 3.

serve que se $\|\mathbf{x}_i\| = 1$ e $\sum_{i=1}^{n} \mathbf{x}_i = 0$, o vetor $\mathrm{X}'\mathbf{v} = (\mathbf{x}'_1\mathbf{v}, \ldots, \mathbf{x}'_p\mathbf{v})'$ é um vetor em que as coordenadas são proporcionais às correlações amostrais. O vetor $\mathbf{v}$ será equiangular se $\mathrm{X}'\mathbf{v} = \mathbf{a} = (a, \ldots, a)' = a\mathbf{1}$. Vamos tomar o vetor $(\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}$. A norma deste vetor, com o produto interno natural do espaço paramétrico $< \mathbf{v}_1, \mathbf{v}_2 >_p = \mathbf{v}'_1 (\mathrm{X}'\mathrm{X}) \mathbf{v}_2$ é

$$\left\| (\mathrm{X}'\mathrm{X})^{-1}\mathbf{1} \right\|_p^2 = \mathbf{1}'(\mathrm{X}'\mathrm{X})^{-1} (\mathrm{X}'\mathrm{X}) (\mathrm{X}'\mathrm{X})^{-1}\mathbf{1} = \mathbf{1}'(\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}.$$

Como se quer vetores unitários, seja $\mathbf{w} = \frac{(\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}}{\left(\mathbf{1}'(\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}\right)^{1/2}}$, e $\mathbf{u} = \mathrm{X}\mathbf{w}$. O vetor $\mathbf{u}$ é unitário pois

$$< \mathbf{u}, \mathbf{u} >= (\mathrm{X}\mathbf{w})' (\mathrm{X}\mathbf{w}) = \mathbf{w}' (\mathrm{X}'\mathrm{X}) \mathbf{w} = 1.$$

O vetor $\mathbf{u}$ é um vetor equiangular em relação a todas as colunas de X pois

$$\mathrm{X}'\mathbf{u} = \mathrm{X}'\mathrm{X}\mathbf{w} = \frac{(\mathrm{X}'\mathrm{X}) (\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}}{\mathbf{1}'(\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}} = \frac{1}{\mathbf{1}'(\mathrm{X}'\mathrm{X})^{-1}\mathbf{1}}\mathbf{1}.$$

Portanto, em termos de complexidade computacional, a obtenção de vetores equiangulares é da mesma complexidade da multiplicação matricial.

De forma geral, o algoritmo é descrito algebricamente pela construção conforme se segue. Seja $\mathcal{A} \subset \{1, 2, \ldots, \mathrm{p}\}$ um subconjunto de indices, os quais indexam as covariáveis. Se $|\mathcal{A}|$ é o número de elementos de $\mathcal{A}$, seja $\mathrm{X}_{\mathcal{A}}$ a matriz $\mathrm{n} \times (|\mathcal{A}|)$ obtida de X tomando-se as colunas correspondentes às covariáveis indexadas por $\mathcal{A}$. Um vetor $\mathbf{u}_{\mathcal{A}}$ será equiangular em relação às colunas de $\mathrm{X}_{\mathcal{A}}$ se vale a relação $\mathrm{X'}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} = \mathrm{a}\mathbf{1}_{\mathcal{A}}$, em que $\mathbf{1}_{\mathcal{A}} = \mathbf{1}_{(|\mathcal{A}|)\times 1} = (1, \ldots, 1)'$. Sejam $\mathrm{G}_{\mathcal{A}} = \mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}$, $\left(\mathbf{1'}_{\mathcal{A}}\mathrm{G^{-1}}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}\right)^{-\frac{1}{2}} = \mathrm{A}_{\mathcal{A}}$, $\mathbf{w}_{\mathcal{A}} = \mathrm{A}_{\mathcal{A}}\mathrm{G^{-1}}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}$ e $\mathbf{u}_{\mathcal{A}} = \mathrm{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}}$. Tem-se então que

$$\begin{aligned}
\mathrm{X'}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} &= \mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}} \\
&= \mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\left(\mathrm{A}_{\mathcal{A}}\mathrm{G^{-1}}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}\right) \\
&= \mathrm{A}_{\mathcal{A}}\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\left(\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)^{-1}\mathbf{1}_{\mathcal{A}}\right) \\
&= \mathrm{A}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}},
\end{aligned}$$

e portanto $\mathbf{u}_{\mathcal{A}}$ é equiangular com os vetores coluna de X definidos pelo subconjunto $\mathcal{A}$. Além disso,

$$\begin{aligned}
\|\mathbf{u}_{\mathcal{A}}\|^2 &= < \mathbf{u}_{\mathcal{A}}, \mathbf{u}_{\mathcal{A}} > \\
&= < \mathrm{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}}, \mathrm{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}} > \\
&= \mathbf{w'}_{\mathcal{A}}\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}} \\
&= \left(\mathrm{A}_{\mathcal{A}}\mathrm{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}\right)'\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\left(\mathrm{A}_{\mathcal{A}}\mathrm{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}\right) \\
&= \mathrm{A}_{\mathcal{A}}^2\mathbf{1'}_{\mathcal{A}}\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)^{-1}\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)^{-1}\mathbf{1}_{\mathcal{A}} \\
&= \mathrm{A}_{\mathcal{A}}^2\mathbf{1'}_{\mathcal{A}}\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)\mathbf{1}_{\mathcal{A}} \\
&= \left[\left(\mathbf{1'}_{\mathcal{A}}\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)^{-1}\mathbf{1}_{\mathcal{A}}\right)^{-\frac{1}{2}}\right]^2\mathbf{1'}_{\mathcal{A}}\left(\mathrm{X'}_{\mathcal{A}}\mathrm{X}_{\mathcal{A}}\right)\mathbf{1}_{\mathcal{A}} \\
&= \mathbf{1}.
\end{aligned}$$

Tem-se portanto um procedimento algébrico para a obtenção de um vetor que seja equiangular à um subconjunto qualquer de colunas da matriz X.

Pode-se agora construir o algoritmo LARS. No primeiro passo considera-se $\hat{\boldsymbol{\mu}}_0 = 0$. O vetor de correlações amostrais é dado por

$$\mathbf{c}(\hat{\boldsymbol{\mu}}_0) = (\mathrm{c}_1(\hat{\boldsymbol{\mu}}_0), \ldots, \mathrm{c}_\mathrm{n}(\hat{\boldsymbol{\mu}}_0)) = \mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}_0) = \mathrm{X}'\mathbf{y}.$$

A ideia é se tomar, assim como no *Step Forward*, a covariável que define a coluna de maior correlação amostral com o vetor de dados $\mathbf{y}$. Seja $\hat{\mathrm{C}} = \max_{\mathrm{j}\in\{1,\ldots,\mathrm{p}\}} \{|\mathrm{c}_\mathrm{j}(\hat{\boldsymbol{\mu}}_0)|\}$. Se $\widetilde{\mathrm{j}}$ é o índice correspondente à covariável de correlação amostral máxima em módulo, então toma-se inicialmente $\mathcal{A} = \left\{\widetilde{\mathrm{j}}\right\}$, $\mathrm{X}_\mathcal{A} = \mathrm{X}_{\widetilde{\mathrm{j}}}$, $\mathrm{G}_\mathcal{A} = \mathrm{X}'_{\widetilde{\mathrm{j}}}\mathrm{X}_{\widetilde{\mathrm{j}}} = 1$, $\mathrm{A}_\mathcal{A} = \left(\mathbf{1}'(\mathrm{X}'_{\widetilde{\mathrm{j}}}\mathrm{X}_{\widetilde{\mathrm{j}}})^{-1}\mathbf{1}\right)^{-\frac{1}{2}} = (\mathbf{1}'\mathbf{1})^{-\frac{1}{2}} = \mathrm{p}^{-\frac{1}{2}}$, $\mathbf{w}_\mathcal{A} = \mathrm{A}_\mathcal{A}(\mathrm{G}_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A} = \frac{1}{\sqrt{\mathrm{p}}}\mathbf{1}_\mathcal{A}$, $\mathbf{u}_\mathcal{A} = \mathrm{X}_\mathcal{A}\mathbf{w}_\mathcal{A} = \mathbf{x}_{\widetilde{\mathrm{j}}}\frac{1}{\sqrt{\mathrm{p}}}\mathbf{1}_\mathcal{A} = \frac{1}{\sqrt{\mathrm{p}}}\mathbf{x}_{\widetilde{\mathrm{j}}}$ e $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\gamma}\mathbf{u}_\mathcal{A} = \hat{\gamma}\frac{1}{\sqrt{\mathrm{p}}}\mathbf{x}_{\widetilde{\mathrm{j}}}$.

O valor de $\hat{\gamma}$ é escolhido da seguinte forma: considere a curva $\mu(\gamma) = \gamma\mathbf{u}_\mathcal{A} = \gamma\frac{1}{\sqrt{\mathrm{p}}}\mathbf{x}_{\widetilde{\mathrm{j}}}$. Tem-se então as curvas de correlação $\mathrm{c}_\mathrm{j}(\gamma) = \mathbf{x}'_\mathrm{j}\left(\mathbf{y} - \gamma\frac{1}{\sqrt{\mathrm{p}}}\mathbf{x}_{\widetilde{\mathrm{j}}}\right)$, para $\mathrm{j} = 1, \ldots, \mathrm{n}$. Para o índice $\widetilde{\mathrm{j}}$, se a correlação é positiva, o valor $\mathrm{c}_{\widetilde{\mathrm{j}}}(\gamma)$ é decrescente, pois

$$\mathrm{c}_{\widetilde{\mathrm{j}}}(\gamma) = \mathbf{x}'_{\widetilde{\mathrm{j}}}\left(\mathbf{y} - \gamma\frac{1}{\sqrt{\mathrm{p}}}\mathbf{x}_{\widetilde{\mathrm{j}}}\right) = \mathbf{x}'_{\widetilde{\mathrm{j}}}\mathbf{y} - \mathbf{x}'_{\widetilde{\mathrm{j}}}\gamma\frac{1}{\sqrt{\mathrm{p}}}\mathbf{x}_{\widetilde{\mathrm{j}}} = \hat{\mathrm{C}} - \frac{\gamma}{\sqrt{\mathrm{p}}}.$$

Toma-se então como $\hat{\gamma}$ o menor valor de γ para o qual existe outra covariável de índice $\widetilde{\mathrm{i}}$ tal que $\mathrm{c}_{\widetilde{\mathrm{i}}}(\gamma) = \mathrm{c}_{\widetilde{\mathrm{j}}}(\gamma)$. Tal procedimento será melhor explicado no passo genérico a seguir.

Suponha agora que se esteja em um passo do algoritmo onde já se definiu um conjunto $\mathcal{A}$ de índices. Tem-se então computados $\hat{\boldsymbol{\mu}}_\mathcal{A}$, $\mathrm{X}_\mathcal{A}$, $\mathrm{A}_\mathcal{A}$, $\mathrm{G}_\mathcal{A}$ e o vetor equiangular $\mathbf{u}_\mathcal{A}$. Vale destacar que quando se expressa $\hat{\boldsymbol{\mu}}_\mathcal{A}$ no estágio em que o conjunto de índices ativo é $\mathcal{A}$, o vetor $\hat{\boldsymbol{\mu}}_\mathcal{A}$ foi obtido no passo anterior e portanto está em um subespaço de dimensão $|\mathcal{A}| - 1$. Assim, o vetor $\mathbf{u}_\mathcal{A}$, que é o vetor equiangular com todos os vetores $s_\mathrm{j}\mathbf{x}_\mathrm{j}$ para $\mathrm{j} \in \mathcal{A}$, não está no mesmo subespaço que $\hat{\boldsymbol{\mu}}_\mathcal{A}$. Neste estágio do algoritmo, o vetor de correlações amostrais é

$\mathbf{c}(\hat{\boldsymbol{\mu}}_{\mathcal{A}})= \mathrm{X}'\,(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{A}})$ e para os índices em $\mathcal{A}$ possuiu o mesmo valor máximo em módulo, que foi denominado $\hat{\mathrm{C}}$, isto é, com $\hat{\mathrm{C}} \geq |\mathrm{c_j}(\hat{\boldsymbol{\mu}}_A)|, \forall \mathrm{j} = 1, \ldots, \mathrm{n}$.

No próximo passo, um novo índice é acrescentado ao conjunto de índice $\mathcal{A}$, definindo o conjunto $\mathcal{A}+$ e o novo vetor de estimativas $\hat{\boldsymbol{\mu}}_{\mathcal{A}+}=\hat{\boldsymbol{\mu}}_{\mathcal{A}}+\hat{\gamma}\mathbf{u}_{\mathcal{A}+}$. O valor $\hat{\gamma}$ é obtido pela relação

$$\hat{\gamma} = \min^{+} \left\{ \frac{\hat{\mathrm{C}} - \hat{\mathrm{c}}_\mathrm{j}}{\mathrm{A}_{\mathcal{A}} - \mathrm{a}_\mathrm{j}}, \frac{\hat{\mathrm{C}}+\hat{\mathrm{c}}_\mathrm{j}}{\mathrm{A}_{\mathcal{A}}+\mathrm{a}_\mathrm{j}} \right\}, \mathrm{j} \notin \mathcal{A},$$

em que $\min^{+}$ é o mínimo tomado apenas para valores positivos. Esta escolha do valor $\hat{\gamma}$ se justifica da forma como se segue. Se $\boldsymbol{\mu}(\gamma) =\hat{\boldsymbol{\mu}}_{\mathcal{A}} + \gamma\mathbf{u}_{\mathcal{A}}$, então o vetor de correlações é $\mathrm{c}(\gamma) = \mathrm{X}'(\mathbf{y} - \boldsymbol{\mu}(\gamma))$, o que implica em

$$\begin{aligned} \mathrm{c_j}(\gamma) &= \mathbf{x}'_\mathrm{j}(\mathbf{y} - \boldsymbol{\mu}(\gamma)) \\ &= \mathbf{x}'_\mathrm{j}(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{A}} - \gamma\mathbf{u}_{\mathcal{A}}) \\ &= \mathbf{x}'_\mathrm{j}(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathcal{A}}) - \gamma\mathbf{x}'_\mathrm{j}\mathbf{u}_{\mathcal{A}}. \end{aligned}$$

No caso em que $\mathrm{j} \in \mathcal{A}$, $\mathbf{x}'_\mathrm{j}(\mathbf{y} - \hat{\boldsymbol{\mu}}_\mathrm{A}) = \hat{\mathrm{C}}$ e $\mathbf{x}'_\mathrm{j}\mathbf{u}_{\mathcal{A}} = \mathrm{a_j} = \mathrm{A}_{\mathcal{A}}$, ou seja,

$$\mathrm{c_j}(\gamma) = \hat{\mathrm{C}} - \gamma\mathrm{a_j}= \hat{\mathrm{C}} - \gamma\mathrm{A}_{\mathcal{A}}.$$

Portanto as correlações correspondentes aos índices em $\mathcal{A}$ decaem de forma linear.

O valor $\hat{\gamma}$ é obtido quando este valor $\mathrm{c} - \gamma\mathrm{A}_{\mathcal{A}}$ se iguala pela primeira vez a $\mathrm{c_j}(\gamma) = \hat{\mathrm{c}}_\mathrm{j} - \gamma\mathrm{a_j}$, para algum j, com $\mathrm{j} \notin \mathcal{A}$, e portanto,

$$\hat{\mathrm{C}} - \hat{\gamma}\mathrm{A}_{\mathcal{A}} = \hat{\mathrm{c}}_\mathrm{j} - \hat{\gamma}\mathrm{a}_j$$

$$\hat{\gamma}\,(\mathrm{A}_{\mathcal{A}} - \mathrm{a_j}) = \hat{\mathrm{C}} - \hat{\mathrm{c}}_\mathrm{j}$$

$$\hat{\gamma} = \frac{\hat{\mathrm{C}} - \hat{\mathrm{c}}_\mathrm{j}}{\mathrm{A}_{\mathcal{A}} - \mathrm{a_j}}.$$

É necessário também considerar o caso $-\mathrm{c_j}(\gamma)$ relativo à $-\mathbf{x}_\mathrm{j}$. Assim, $\mathrm{a_j} = \mathbf{x}'_\mathrm{j}\boldsymbol{\mu}_{\mathcal{A}}$

troca de sinal e

$$\hat{\mathrm{C}} - \hat{\gamma}\mathrm{A}_{\mathcal{A}} = -\hat{\mathrm{c}}_{\mathrm{j}} + \hat{\gamma}\mathrm{a}_{\mathrm{j}}$$

$$\hat{\gamma} = \frac{\hat{\mathrm{C}} + \hat{\mathrm{c}}_{\mathrm{j}}}{\mathrm{A}_{\mathcal{A}} + \mathrm{a}_{\mathrm{j}}}.$$

Portanto, $\hat{\gamma}$ é o menor valor tal que um novo índice $\hat{\mathrm{j}}$ é incluído em $\mathcal{A}$, $\mathcal{A}+ = \mathcal{A} \cup \{\hat{\mathrm{j}}\}$ de tal forma que a nova correlação máxima em valor absoluto é $\hat{\mathrm{C}}_{+} = \hat{\mathrm{C}} - \hat{\gamma}\mathrm{A}_{\mathcal{A}}$.

6.3 O algoritmo LARS aplicado à delineamentos ortogonais

Para o caso em que $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, com $\mathrm{X}'\mathrm{X} = \mathrm{I}$, o algoritmo LARS fica bastante simplificado. Neste caso, as colunas da matriz X formam em $\mathbb{R}^{\mathrm{n}}$ um conjunto com p vetores ortonormais (Figura 49).

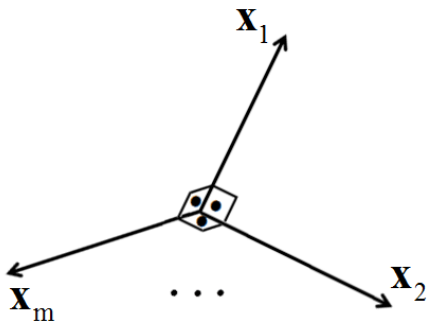


Figura 49 Vetores coluna para X ortonormais.

Por meio de uma transformação linear ortogonal A de $\mathbb{R}^{\mathrm{n}}$, isto é, uma transformação dos dados, tem-se $\mathbb{R}^{\mathrm{p}} \overset{\mathrm{X}}{\longmapsto} \mathbb{R}^{\mathrm{n}} \overset{\mathrm{A}}{\longmapsto} \mathbb{R}^{\mathrm{n}}$, $\mathrm{A} \cdot \mathrm{X} : \mathbb{R}^{\mathrm{p}} \mapsto \mathbb{R}^{\mathrm{n}}$ tal que $\mathrm{A} \cdot \mathrm{X}(\mathbf{e}_{\mathrm{i}}) = \mathbf{e}_{\mathrm{i}}$ e portanto a imagem de $\mathrm{A} \cdot \mathrm{X}$ é a inclusão canônica de $\mathbb{R}^{\mathrm{p}}$ em $\mathbb{R}^{\mathrm{n}}$. Com esta observação pode-se supor, sem perda de generalidade, que as colunas da matriz X são da forma $\mathbf{x}_{\mathrm{i}} = \mathbf{e}_{\mathrm{i}}$, $\mathrm{i} = 1, \ldots, \mathrm{p}$. O vetor de dados $\mathbf{y}' = (\mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{n}})$ se projeta em $\mathbb{R}^{\mathrm{p}}$ como $(\mathrm{P}_{\mathbb{R}^{\mathrm{p}}}\mathbf{y})' = (\mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{p}}, 0, \ldots, 0)$, e portanto

os valores $\mathrm{y_{p+1}}, \mathrm{y_{p+2}}, \ldots, \mathrm{y_n}$ não afetarão o processo de estimação e podem ser considerados nulos. Em outras palavras, todo o processo de estimação ocorre em $\mathbb{R}^{\mathrm{p}}$ e a matriz X é a identidade. Desta forma tem-se para um conjunto de índices $\mathcal{A}$ em um passo do algoritmo que $\mathrm{G}_{\mathcal{A}} = \mathrm{X}'_{\mathcal{A}}\mathrm{X}_{\mathcal{A}} = \mathrm{I}_{\mathcal{A}}$. Também se tem que

$$\mathrm{A}_{\mathcal{A}} = \left(\mathbf{1}'_{\mathcal{A}}\mathrm{G}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}\right)^{-1/2} = \left(\mathbf{1}'_{\mathcal{A}}\mathrm{I}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}\right)^{-1/2} = \left(\mathbf{1}'_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}\right)^{-1/2} = \frac{1}{\sqrt{|\mathcal{A}|}},$$

$$\mathbf{w}_{\mathcal{A}} = \mathrm{A}_{\mathcal{A}}\mathrm{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}} = \frac{1}{\sqrt{|\mathcal{A}|}}\mathrm{I}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}} = \frac{1}{\sqrt{|\mathcal{A}|}}\mathrm{I}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}} = \frac{1}{\sqrt{|\mathcal{A}|}}\mathbf{1}_{\mathcal{A}},$$

e

$$\mathbf{u}_{\mathcal{A}} = \mathrm{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}} = \frac{1}{\sqrt{|\mathcal{A}|}}\mathrm{X}_{\mathcal{A}}\mathbf{1}_{\mathcal{A}} = \frac{1}{\sqrt{|\mathcal{A}|}}\mathbf{1}_{\mathcal{A}}.$$

O vetor de correlações é o próprio vetor $\mathbf{y}$, uma vez que $\mathrm{X}'\mathbf{y} = \mathrm{I}'\mathbf{y} = \mathbf{y}$. Como temos que tomar no primeiro passo o índice da covariável de maior correlação com $\mathbf{y}$, basta então tomar o índice relativo a maior coordenada do vetor $\mathbf{y}$, a menos do sinal. É conveniente utilizar a notação de estatística de ordem. Considere então as estatísticas de ordem tomadas em relação aos valores absolutos das coordenadas. Para simplificar, será suposto que empates não ocorrem. Assim, $|\mathbf{y}|_{(1)} > |\mathbf{y}|_{(2)} > \cdots > |\mathbf{y}|_{(\mathrm{p})} \geq 0$. Portanto no primeiro passo do algoritmo é escolhida a covariável de índice $\mathrm{j}(1)$ relativo à coordenada $|\mathbf{y}|_{(1)}$, isto é, $\mathcal{A} = \{\mathrm{j}(1)\}$. Logo,

$$\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\gamma}_1\mathbf{u}_{\mathcal{A}} = \mathbf{0} + \hat{\gamma}\frac{1}{\sqrt{|1|}}\mathbf{1}_{\mathcal{A}} = \hat{\gamma}_1\mathbf{e}_{\mathrm{j}(1)}.$$

O valor de $\hat{\gamma}_1$ é obtido tal que este seja o menor valor de tal forma que a correlação entre $\mathbf{x}_{\mathrm{j}(1)}$ e $\mathbf{y} - \hat{\gamma}_1\mathbf{x}_{\mathrm{j}(1)}$ seja igual à correlação de outra covariável $\mathbf{x}_{\mathrm{s}}$, isto é,

$$\mathbf{x}'_{\mathrm{j}(1)}\left(\mathbf{y} - \hat{\gamma}_1\mathbf{x}_{\mathrm{j}(1)}\right) = |\mathbf{y}|_{(1)} - \hat{\gamma}_1 = \mathbf{x}'_{\mathrm{s}}\left(\mathbf{y} - \hat{\gamma}_1\mathbf{x}_{\mathrm{j}(1)}\right) = \mathbf{x}'_{\mathrm{s}}\mathbf{y} = \mathbf{y}_{\mathrm{s}}.$$

Logo, $|\mathbf{y}|_{(1)} - \hat{\gamma}_1 = \mathbf{y}_{\mathrm{s}} \Rightarrow |\mathbf{y}|_{(1)} - \mathbf{y}_{\mathrm{s}} = \hat{\gamma}_1$. O menor valor ocorre para o maior valor possível de $\mathbf{y}_{\mathrm{s}}$ que é $|\mathbf{y}|_{(2)}$. Portanto $\hat{\gamma}_1 = |\mathbf{y}|_{(1)} - |\mathbf{y}|_{(2)}$ e o índice

$\mathrm{j}(2)$ relativo a $|\mathbf{y}|_{(2)}$ é adicionado ao conjunto de índices ativos obtendo-se $\mathcal{A} = \{\mathrm{j}(1), \mathrm{j}(2)\}$. Para se evitar o emprego de indução um passo a mais no algoritmo será desenvolvido deixando claro como é a forma geral.

O vetor unitário equiangular em relação a $\mathbf{e}_{\mathrm{j}(1)}$ e $\mathbf{e}_{\mathrm{j}(2)}$ é

$$\mathbf{u}_{\mathcal{A}} = \frac{1}{\sqrt{2}}\mathbf{1}_{\mathcal{A}} = \frac{1}{\sqrt{2}}\left(0, \ldots, 1, 0, \ldots, 1, 0, \ldots, 0\right),$$

com 1 nas $\mathrm{j}(1)$ e $\mathrm{j}(2)$ coordenadas. A estimativa LARS é atualizada para $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\gamma}_2 \frac{\mathbf{1}_{\mathcal{A}}}{\sqrt{2}}$. Novamente o valor de $\hat{\gamma}_2$ é o menor valor para o qual uma outra covariável $\mathbf{x}_{\mathrm{s}}$ tem a mesma correlação de $\mathbf{x}_{\mathrm{j}(1)}$ e $\mathbf{x}_{\mathrm{j}(2)}$ com $\mathbf{y} - \hat{\boldsymbol{\mu}}_2$. Como $\mathbf{x}_{\mathrm{j}(1)} = \mathbf{e}_{\mathrm{j}(1)}$, $\mathbf{x}_{\mathrm{j}(2)} = \mathbf{e}_{\mathrm{j}(2)}$ e $\mathbf{x}_{\mathrm{s}} = \mathbf{e}_{\mathrm{s}}$, tem-se

$$\mathbf{e}'_{\mathrm{j}(1)}(\mathbf{y} - \hat{\boldsymbol{\mu}}_2) = \mathbf{e}'_{\mathrm{j}(2)}(\mathbf{y} - \hat{\boldsymbol{\mu}}_2) = \mathbf{e}'_{\mathrm{s}}(\mathbf{y} - \hat{\boldsymbol{\mu}}_2)$$

$$\Rightarrow \mathbf{e}'_{\mathrm{j}(1)}\mathbf{y} - \mathbf{e}'_{\mathrm{j}(1)}\hat{\boldsymbol{\mu}}_2 = \mathbf{e}'_{\mathrm{j}(2)}\mathbf{y} - \mathbf{e}'_{\mathrm{j}(2)}\hat{\boldsymbol{\mu}}_2 = \mathbf{e}'_{\mathrm{s}}\mathbf{y} = \mathbf{y}_{\mathrm{s}}$$

$$\Rightarrow |\mathbf{y}|_{(1)} - \mathbf{e}'_{\mathrm{j}(1)}\left(\hat{\gamma}_1\mathbf{e}_{\mathrm{j}(1)} + \hat{\gamma}_2\frac{\mathbf{1}_{\mathcal{A}}}{\sqrt{2}}\right) = |\mathbf{y}|_{(2)} - \mathbf{e}'_{\mathrm{j}(2)}\left(\hat{\gamma}_1\mathbf{e}_{\mathrm{j}(1)} + \hat{\gamma}_2\frac{\mathbf{1}_{\mathcal{A}}}{\sqrt{2}}\right) = \mathbf{y}_{\mathrm{s}}$$

$$\Rightarrow |\mathbf{y}|_{(1)} - \hat{\gamma}_1 - \frac{\hat{\gamma}_2}{\sqrt{2}} = |\mathbf{y}|_{(2)} - \frac{\hat{\gamma}_2}{\sqrt{2}} = \mathbf{y}_{\mathrm{s}}$$

$$\Rightarrow |\mathbf{y}|_{(1)} - \left(|\mathbf{y}|_{(1)} - |\mathbf{y}|_{(2)}\right) - \frac{\hat{\gamma}_2}{\sqrt{2}} = |\mathbf{y}|_{(2)} - \frac{\hat{\gamma}_2}{\sqrt{2}} = \mathbf{y}_{\mathrm{s}}$$

$$\Rightarrow |\mathbf{y}|_{(2)} - \frac{\hat{\gamma}_2}{\sqrt{2}} = \mathbf{y}_{\mathrm{s}}$$

$$\Rightarrow |\mathbf{y}|_{(2)} - \mathbf{y}_{\mathrm{s}} = \frac{\hat{\gamma}_2}{\sqrt{2}}.$$

Portanto o menor valor de $\hat{\gamma}_2$ ocorre quando $\mathbf{y}_{\mathrm{s}}$ é a coordenada relativa a $|\mathbf{y}|_{(3)}$ e se tem $\hat{\gamma}_2 = \sqrt{2}\left(|\mathbf{y}|_{(2)} - |\mathbf{y}|_{(3)}\right)$. Desta forma, fica claro como evolui o algoritmo. Observe que

$$\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\gamma}_2\mathbf{u}_1 = \hat{\gamma}_1\mathbf{e}_{\mathrm{j}(1)} + \hat{\gamma}_2\mathbf{1}_{\mathcal{A}}$$

$$= \left(|\mathbf{y}|_{(1)} - |\mathbf{y}|_{(2)}\right) \mathbf{e}_{\mathrm{j}(1)} + \frac{1}{\sqrt{2}} (0, \ldots, \hat{\gamma}_2, 0, \ldots, \hat{\gamma}_2, 0, \ldots, 0)$$

$$= \left(|\mathbf{y}|_{(1)} - |\mathbf{y}|_{(2)}\right) \mathbf{e}_{\mathrm{j}(1)} + \hat{\gamma}_2 \frac{\mathbf{e}_{\mathrm{j}(1)}}{\sqrt{2}} + \hat{\gamma}_2 \frac{\mathbf{e}_{\mathrm{j}(2)}}{\sqrt{2}}$$

$$= \left(|\mathbf{y}|_{(1)} - |\mathbf{y}|_{(2)}\right) \mathbf{e}_{\mathrm{j}(1)} + \left(|\mathbf{y}|_{(2)} - |\mathbf{y}|_{(3)}\right) \mathbf{e}_{\mathrm{j}(1)} + \left(|\mathbf{y}|_{(3)} - |\mathbf{y}|_{(2)}\right) \mathbf{e}_{\mathrm{j}(2)}$$

$$= \left(|\mathbf{y}|_{(1)} - |\mathbf{y}|_{(3)}\right) \mathbf{e}_{\mathrm{j}(1)} + \left(|\mathbf{y}|_{(2)} - |\mathbf{y}|_{(3)}\right) \mathbf{e}_{\mathrm{j}(2)}.$$

Como $\mathbf{y}_{\mathrm{j}(1)} = |\mathbf{y}|_{(1)}$ e $\mathbf{y}_{\mathrm{j}(2)} = |\mathbf{y}|_{(2)}$, então

$$\hat{\boldsymbol{\mu}}_2 = \left(\mathbf{y}_{\mathrm{j}(1)} - |\mathbf{y}|_{(3)}\right) \mathbf{e}_{\mathrm{j}(1)} + \left(\mathbf{y}_{\mathrm{j}(2)} - |\mathbf{y}|_{(3)}\right) \mathbf{e}_{\mathrm{j}(2)}.$$

Portanto, fica demonstrado o lema:

Lema 1. *Para um delineamento ortogonal com* $\mathbf{x}_{\mathrm{j}} = \mathbf{e}_{\mathrm{j}}$, $\mathrm{j} = 1, \ldots, \mathrm{n}$*, a k-ésima estimativa LARS* $(0 \leq \mathrm{k} \leq \mathrm{n})$ *é dada por*

$$(\hat{\boldsymbol{\mu}}_{\mathrm{k}}(\mathbf{y}))_{\mathrm{i}} = \begin{cases} \mathrm{y_i} - |\mathbf{y}|_{(\mathrm{k}+1)} & se\ \mathrm{y_i} > |\mathbf{y}|_{(\mathrm{k}+1)} \\ 0 & se\ \ |\mathbf{y}| \leq |\mathbf{y}|_{(\mathrm{k}+1)} \\ \mathrm{y_i} + |\mathbf{y}|_{(\mathrm{k}+1)} & se\ \mathrm{y_i} < -|\mathbf{y}|_{(\mathrm{k}+1)}. \end{cases}$$

6.4 Relação entre LARS e OLS

É possível se dar uma descrição do algoritmo LARS em termos de estimadores de quadrados mínimos. Suponha que o algoritmo tenha completado o $\mathrm{k}-1$ passo e portanto com a estimativa $\hat{\boldsymbol{\mu}}_{\mathrm{k}-1}$ e iniciando o k-ésimo passo. O conjunto de índices ativo é $\mathcal{A}_{\mathrm{k}}$ (foi acrescentado o índice k para numerar o passo do algoritmo). Calcula-se então $\mathrm{X_k}$, $\mathrm{G_k}$, $\mathrm{A_k}$ e $\boldsymbol{\mu}_{\mathrm{k}}$ (o sub-índice $\mathcal{A}$ foi substituído por k). Seja $\bar{\mathbf{y}}_{\mathrm{k}}$ a projeção de $\mathbf{y}$ no subespaço gerado pelas colunas de X indexadas por $\mathcal{A}_{\mathrm{k}}$, $(\mathcal{L}(\mathrm{X_k}))$, isto é, $\bar{\mathbf{y}}_{\mathrm{k}} = \mathrm{X_k}(\mathrm{X'_k X_k})^{-1}\mathrm{X'_k}\mathbf{y}$ e portanto $\bar{\mathbf{y}}_{\mathrm{k}}$ define o estimador de quadrados mínimos quando se considera apenas as covariáveis definidas por $\mathcal{A}_{\mathrm{k}}$. Observe que $\hat{\boldsymbol{\mu}}_{\mathrm{k}-1} \in \mathcal{L}(\mathrm{X_{k-1}})$. Pode-se então escrever, observando que

$\mathrm{X_k G_k^{-1} X'_k \hat{\boldsymbol{\mu}}_{k-1}} = \hat{\boldsymbol{\mu}}_{\mathrm{k}-1}$ pois $\mathrm{X_k G_k^{-1} X'_k} = \mathrm{X_k (X'_k X_k)^{-1} X'_k}$ é um projetor no espaço $\mathcal{L}(\mathrm{X})$ e $\hat{\boldsymbol{\mu}}_{\mathrm{k}-1} \in \mathcal{L}(\mathrm{X}_{\mathrm{k}-1})$, que

$$\begin{aligned}
\bar{\mathbf{y}}_{\mathrm{k}} &= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \bar{\mathbf{y}}_{\mathrm{k}} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} \\
&= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \mathrm{X_k G_k^{-1} X'_k} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right) \\
&= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \mathrm{X_k G_k^{-1}} \hat{\mathrm{C}} \mathbf{1}_{\mathcal{A}} \\
&= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \frac{1}{\mathrm{A}_{\mathcal{A}}} \mathrm{X_k A_{\mathcal{A}} G_k^{-1}} \hat{\mathrm{C}} \mathbf{1}_{\mathcal{A}} \\
&= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \frac{1}{\mathrm{A}_{\mathcal{A}}} \hat{\mathrm{C}} \mathrm{X_k A_{\mathcal{A}} G_k^{-1}} \mathbf{1}_{\mathcal{A}} \\
&= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \frac{\hat{\mathrm{C}}}{\mathrm{A}_{\mathcal{A}}} \mathrm{X_k} \mathbf{w}_{\mathcal{A}} \\
&= \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \frac{\hat{\mathrm{C}}}{\mathrm{A}_{\mathcal{A}}} \mathbf{u}_{\mathrm{k}},
\end{aligned}$$

e portanto,

$$\left\| \bar{\mathbf{y}}_{\mathrm{k}} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} \right\| = \left\| \frac{\hat{\mathrm{C}}}{\mathrm{A}_{\mathcal{A}}} \mathbf{u}_{\mathrm{k}} \right\| = \frac{\hat{\mathrm{C}}}{\mathrm{A}_{\mathcal{A}}} = \bar{\gamma}.$$

Como $\hat{\boldsymbol{\mu}}_{\mathrm{k}} = \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \hat{\gamma} \mathbf{u}_{\mathrm{k}}$ e $\bar{\mathbf{y}}_{\mathrm{k}} = \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \bar{\gamma} \mathbf{u}_{\mathrm{k}}$, os vetores $\hat{\boldsymbol{\mu}}_{\mathrm{k}} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1}$ e $\bar{\mathbf{y}}_{\mathrm{k}} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1}$ são colineares. Prova-se que $\hat{\gamma} < \bar{\gamma}$ (Figura 50). Ao se atingir $\mathcal{A}_{\mathrm{n}} = \{1, \ldots, \mathrm{n}\}$, pela definição do algoritmo $\hat{\boldsymbol{\mu}}_{\mathrm{n}} = \bar{\mathbf{y}}_{\mathrm{n}} = \hat{\boldsymbol{\mu}}_{\mathrm{ols}}$.

6.5 Relação entre o algoritmo LARS e o método LASSO

Uma propriedade surpreendente do algoritmo LARS é que com pequenas modificações o algoritmo é capaz de computar as estimativas obtidas pelo método LASSO. Observa-se que em relação ao parâmetro t do método LASSO, as estimativas $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}(\mathrm{t})$ definem um caminho que se inicia em $\boldsymbol{\beta} = 0$ e termina no estimador de quadrados mínimos $\hat{\boldsymbol{\beta}}^{\mathrm{ols}}$. Tem-se que este caminho no espaço de parâmetros define também um caminho no espaço de dados $\hat{\boldsymbol{\mu}}(\mathrm{t}) = \mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}(\mathrm{t})$. De forma semelhante, o algoritmo LARS define um caminho no espaço de dados formado por

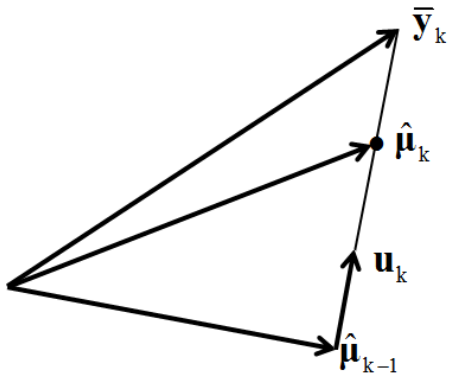


Figura 50 Relação LARS e OLS.

segmentos de retas, no k-ésimo passo dado por $\hat{\boldsymbol{\mu}}_{\mathrm{k}}(\gamma) = \hat{\boldsymbol{\mu}}_{\mathrm{k}-1} + \gamma\left(\overline{\mathbf{y}}_{\mathrm{k}} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right)$. Com as devidas modificações, as duas curvas serão as mesmas. As modificações são introduzidas a partir de uma longa sequência de lemas. Estes são bastante técnicos e neste trabalho serão abordados apenas alguns deles. Como referência completa, tem-se o artigo original (EFRON et al., 2004). Os três primeiros lemas são propriedades gerais do LARS.

Lema 2. *Suponha que ao fim do* $\mathrm{k}-1$ *passo do algoritmo LARS a covariável definida digamos pela coluna* $\mathbf{x}_{\mathrm{k}}$ *seja adicionada ao conjunto de índices* $\mathcal{A}_{\mathrm{k}-1}$ *(sem perda de generalidade as covariáveis serão enumeradas por* $1, 2, 3, \ldots$ *na ordem em que estas são incluídas nos passos consecutivos do algoritmo). O vetor* $\mathbf{w}_{\mathrm{k}} = \mathrm{A}_{\mathrm{k}}(G_{\mathrm{k}})^{-1}\mathbf{1}_{\mathrm{k}}$ *que define o vetor equiangular* $\mathbf{u}_{\mathrm{k}} = \mathrm{X}_{\mathrm{k}}\mathbf{w}_{\mathrm{k}}$ *possui sua k-ésima componente* $\mathbf{w}_{\mathrm{kk}}$ *com o mesmo sinal da correlação* $\mathrm{c}_{\mathrm{kk}} = \mathbf{x}'_{\mathrm{k}}\left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right)$. *Além disso, o vetor* $\hat{\boldsymbol{\beta}}_{\mathrm{k}}$*, com* $\hat{\boldsymbol{\mu}}_{\mathrm{k}} = \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{k}}$*, tem sua k-ésima componente* $\hat{\boldsymbol{\beta}}_{\mathrm{kk}}$ *também com o mesmo sinal.*

Demonstração. Como $\mathbf{X'}_{\mathrm{k}}\left(\mathbf{y}-\hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right)=\hat{\mathrm{C}}_{\mathrm{k}}\mathbf{1}_{\mathrm{k}}$, então

$$\begin{aligned}\mathbf{w}_{\mathrm{k}} &= \mathrm{A}_{\mathrm{k}}(G_{\mathrm{k}})^{-1}\mathbf{1}_{\mathrm{k}}\\ &= \mathrm{A}_{\mathrm{k}}(G_{\mathrm{k}})^{-1}(\hat{\mathrm{C}}_{\mathrm{k}})^{-1}\mathrm{X'}_{\mathrm{k}}\left(\mathbf{y}-\hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right)\\ &= \mathrm{A}_{\mathrm{k}}(\hat{\mathrm{C}}_{\mathrm{k}})^{-1}\left(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}}\right)^{-1}\mathrm{X'}_{\mathrm{k}}\left(\mathbf{y}-\hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right)\\ &= \mathrm{A}_{\mathrm{k}}(\hat{\mathrm{C}}_{\mathrm{k}})^{-1}\mathbf{w}^{*}_{\mathrm{k}}.\end{aligned}$$

Tem-se que $\mathrm{X'}_{\mathrm{k}}\left(\mathbf{y}-\bar{\mathbf{y}}_{\mathrm{k}-1}\right)=\mathrm{X'}_{\mathrm{k}}\left(\bar{\mathbf{y}}_{\mathrm{k}}-\bar{\mathbf{y}}_{\mathrm{k}-1}\right)=(\underbrace{0}_{\mathrm{k}-1},\underbrace{\delta}_{1})'$ pois $\bar{\mathbf{y}}_{\mathrm{k}}$ e $\bar{\mathbf{y}}_{\mathrm{k}-1}$ possuem as primeiras $\mathrm{k}-1$ coordenadas iguais, isto é, $\mathrm{X'}_{\mathrm{k}-1}\left(\bar{\mathbf{y}}_{\mathrm{k}}-\bar{\mathbf{y}}_{\mathrm{k}-1}\right)=0$. Considere as curvas $\mathrm{c_i}(\gamma)=\mathbf{x'}_{\mathrm{i}}\left(\mathbf{y}-\gamma\mathbf{u}_{\mathrm{k}-1}\right)$, $\mathrm{i}=1,\ldots,\mathrm{k}$. Pelo fato de o vetor $\mathbf{u}_{\mathrm{k}-1}$ ser equiangular com as covariáveis $\mathbf{x}_{\mathrm{i}}$, $\mathrm{i}=1,\ldots,\mathrm{k}-1$, tem-se

$$\mathrm{c_k}(\gamma)\begin{cases}<\mathrm{c_j}(\gamma) & \gamma<\hat{\gamma}_{\mathrm{k}-1}\\ =\mathrm{c_j}(\gamma)=\hat{\mathrm{c}}_{\mathrm{k}} & \gamma=\hat{\gamma}_{\mathrm{k}-1}\\ >\mathrm{c_j}(\gamma) & \hat{\gamma}_{\mathrm{k}-1}<\gamma<\bar{\gamma}_{\mathrm{k}-1}\end{cases}\qquad \mathrm{j}<\mathrm{k}.$$

Portanto,

$$\begin{aligned}\mathbf{w}^{*}_{\mathrm{k}} &= \left(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}}\right)^{-1}\mathrm{X'}_{\mathrm{k}}\left(\bar{\mathbf{y}}_{\mathrm{k}}-\bar{\mathbf{y}}_{\mathrm{k}-1}+\bar{\mathbf{y}}_{\mathrm{k}-1}-\hat{\mu}_{\mathrm{k}-1}\right)\\ &= \left(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}}\right)^{-1}\begin{pmatrix}0\\ \delta\end{pmatrix}+\left(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}}\right)^{-1}\left(\bar{\mathbf{y}}_{\mathrm{k}-1}-\hat{\boldsymbol{\mu}}_{\mathrm{k}-1}\right)\\ &= \left(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}}\right)^{-1}\begin{pmatrix}0\\ \delta\end{pmatrix}+0.\end{aligned}$$

$(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}})$ é positiva definida e portanto $(0,\delta)(\mathrm{X'}_{\mathrm{k}}\mathrm{X}_{\mathrm{k}})^{-1}\begin{pmatrix}0\\ \delta\end{pmatrix}=\mathrm{a_{kk}}\delta^{2}>0$, em

que $\mathrm{a_{kk}}$ é a entrada correspondente de $(\mathrm{X'_k X_k})^{-1} \Rightarrow \mathrm{a_{kk}} > 0$. Assim,

$$\mathbf{w}_\mathrm{k}^* = \left(\mathrm{X'_k X_k}\right)^{-1} \begin{pmatrix} 0 \\ \delta \end{pmatrix} = \begin{pmatrix} * \\ * \\ \vdots \\ \mathrm{a_{kk}}\delta \end{pmatrix}$$

e $\mathrm{a_{kk}}\delta > 0$. Portanto, a k-ésima componente de $\mathbf{w}_\mathrm{k} = \mathrm{A_k}\hat{\mathrm{c}}_\mathrm{k}^{-1}\mathbf{w}_\mathrm{k}^*$ tem o mesmo sinal de $\hat{\mathrm{c}}_\mathrm{k}^{-1} = \mathrm{c_{kk}}$. □

O segundo lema interpreta a quantidade $\mathrm{A}_\mathcal{A} = \left(\mathbf{1}'_\mathcal{A}(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A}\right)^{-\frac{1}{2}}$. Seja $\mathcal{S}_\mathcal{A}$ o simplexo estendido

$$\mathcal{S}_\mathcal{A} = \left\{ \mathbf{v} = \sum_{\mathrm{j}\in\mathcal{A}} s_\mathrm{j}\mathbf{x}_\mathrm{j}\mathrm{P_j} \;;\; \sum_{\mathrm{j}\in\mathcal{A}} \mathrm{P_j} = 1 \right\},$$

em que extendido significa que $\mathrm{P_j}$ pode ser negativo. O simplexo estendido $\mathcal{S}_\mathcal{A}$ está contido em um hiperplano que passa pelos pontos definidos pelos vetores $\{s_\mathrm{j}\mathbf{x}_\mathrm{j},\, \mathrm{j} \in \mathcal{A}\}$. Recordando que $s_\mathrm{j} = \pm 1$ definidos de forma que $s_\mathrm{j}\mathbf{x}_\mathrm{j}$ tenha um ângulo menor do que 90° com o vetor de dados $\mathbf{y}$, tal situação é representada na Figura 51.

Lema 3. *O ponto de $\mathcal{S}_\mathcal{A}$ mais próximo à origem é definido pelo vetor $\mathbf{v}_\mathcal{A} = \mathrm{A}_\mathcal{A}\mathbf{u}_\mathcal{A} = \mathrm{A}_\mathcal{A}\mathrm{X}_\mathcal{A}\mathbf{w}_\mathcal{A}$ em que $\mathbf{w}_\mathcal{A} = \mathrm{A}_\mathcal{A}(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A}$ e $\|\mathbf{v}_\mathcal{A}\| = \mathrm{A}_\mathcal{A}$. Se $\mathcal{A} \subseteq \mathcal{B}$ então $\mathrm{A}_\mathcal{A} \geq \mathrm{A}_\mathcal{B}$, sendo que o maior valor possível é $\mathrm{A}_\mathcal{A} = 1$ para um $\mathcal{A}$ com apenas um ponto.*

Demonstração. Considere o vetor $\mathbf{v} = \mathrm{X}_\mathcal{A}\mathrm{P}$, em que $\mathrm{P} = (\mathrm{p}_1, \ldots, \mathrm{p}_{|\mathcal{A}|})$ e portanto

$$\|\mathbf{v}\|^2 = (\mathrm{X}_\mathcal{A}\mathrm{P})'\,(\mathrm{X}_\mathcal{A}\mathrm{P}) = \mathrm{P'X'}_\mathcal{A}\mathrm{X}_\mathcal{A}\mathrm{P} = \mathrm{P'}G_\mathcal{A}\mathrm{P}.$$

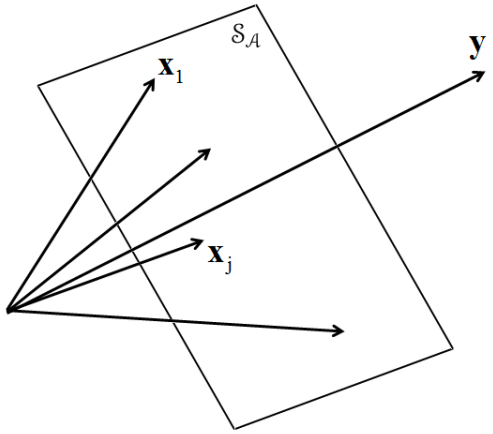


Figura 51 Geometria do conjunto $\mathcal{S}_\mathcal{A}$

Tem-se o problema de minimizar $\mathrm{P}'G_\mathcal{A}\mathrm{P}$ sujeito à restrição $\sum_{\mathrm{j}\in\mathcal{A}} \mathrm{P}_\mathrm{j} = 1$ que pode ser descrito como $\mathbf{1}'_\mathcal{A}\mathrm{P} = 1$. Pelo método dos multiplicadores de Lagrange, tem-se a lagrangiana $\mathrm{P}'G_\mathcal{A}\mathrm{P} - \lambda\left(\mathbf{1}'_\mathcal{A}\mathrm{P} - 1\right)$. Derivando em relação à P tem-se que o mínimo ocorre para $\mathrm{P}_\mathcal{A} = \lambda(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A}$. Como a soma das componentes de $\mathrm{P}_\mathcal{A}$ é 1, tem-se então

$$1 = \mathbf{1}'_\mathcal{A}\mathrm{P}_\mathcal{A} = \mathbf{1}'_\mathcal{A}\lambda(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A} = \lambda\mathbf{1}'_\mathcal{A}(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A} = \lambda\mathcal{A}^{-2} \Rightarrow \lambda = \mathrm{A}^2_\mathcal{A},$$

o que implica em

$$\mathrm{P}_\mathcal{A} = \mathrm{A}^2_\mathcal{A}(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A} = \mathrm{A}_\mathcal{A}\mathrm{A}_\mathcal{A}(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A} = \mathrm{A}_\mathcal{A}\mathbf{w}_\mathcal{A}.$$

Logo, como $\mathbf{v}_\mathcal{A} = \mathrm{X}_\mathcal{A}\mathrm{P}_\mathcal{A} = \sum_{\mathrm{j}\in\mathcal{A}} s_\mathrm{j}\mathbf{x}'_\mathrm{j}\mathrm{P}_\mathrm{j}$, tem-se $\mathbf{v}_\mathcal{A} \in \mathcal{S}_\mathcal{A}$, e

$$\|\mathbf{v}_\mathcal{A}\|^2 = \mathrm{P}'_\mathcal{A}(G_\mathcal{A})^{-1}\mathrm{P}_\mathcal{A} = \mathrm{A}_\mathcal{A}\mathbf{1}_\mathcal{A}(G_\mathcal{A})^{-1}\mathbf{1}_\mathcal{A} = \mathrm{A}^2_\mathcal{A}.$$

Se $\mathcal{A} \subseteq \mathcal{B} \Rightarrow \mathcal{S}_\mathcal{A} \subseteq \mathcal{S}_\mathcal{B}$, o ponto mais próximo à origem de $\mathcal{S}_\mathcal{B}$ deve ser menor do que o ponto mais próximo de $\mathcal{S}_\mathcal{A}$ implicando então que $\mathrm{A}_\mathcal{A} \geq \mathrm{A}_\mathcal{B}$. □

Como o algoritmo LARS evolui por segmentos de retas, as alterações são realizadas em cada segmento linear.

Considere o caminho $\boldsymbol{\beta}(\gamma) = \boldsymbol{\beta} + \gamma\mathbf{d}$, em que $\boldsymbol{\beta}$ e $\mathbf{d}$ são vetores em $\mathbb{R}^{\mathrm{p}}$ e $S(\gamma) = \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}(\gamma)\|^2$. Segue então o lema

Lema 4. *Se* $\mathbf{c} = \mathrm{X}'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})$ *é o vetor de correlações no passo em que* $\boldsymbol{\mu} = \mathrm{X}\boldsymbol{\beta}$*, então* $S(\gamma) - S(0) = -2(\mathbf{c}'\mathbf{d})\gamma + \mathbf{d}'(\mathrm{X}'\mathrm{X})\mathbf{d}\gamma^2$.

Demonstração.

$$
\begin{aligned}
S(\gamma) - S(0) &= (\mathbf{y} - \mathrm{X}\boldsymbol{\beta}(\gamma))'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}(\gamma)) - (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \\
&= (\mathbf{y} - \mathrm{X}(\boldsymbol{\beta} + \gamma\mathbf{d}))'(\mathbf{y} - \mathrm{X}(\boldsymbol{\beta} + \gamma\mathbf{d})) - (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \\
&= (\mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\gamma\mathbf{d})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\gamma\mathbf{d}) - (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \\
&= \left((\mathbf{y} - \mathrm{X}\boldsymbol{\beta})' - (\mathrm{X}\gamma\mathbf{d})'\right)((\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) - (\mathrm{X}\gamma\mathbf{d})) \\
&\quad -(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \\
&= (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) - (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathrm{X}\gamma\mathbf{d}) \\
&\quad -(\mathrm{X}\gamma\mathbf{d})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) + (\mathrm{X}\gamma\mathbf{d})'(\mathrm{X}\gamma\mathbf{d}) - (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) \\
&= -(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})'(\mathrm{X}\gamma\mathbf{d}) - (\mathrm{X}\gamma\mathbf{d})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) + (\mathrm{X}\gamma\mathbf{d})'(\mathrm{X}\gamma\mathbf{d}) \\
&= -(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})(\mathrm{X}\gamma\mathbf{d})' - (\mathrm{X}\gamma\mathbf{d})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) + (\mathrm{X}\gamma\mathbf{d})'(\mathrm{X}\gamma\mathbf{d}) \\
&= -2(\mathrm{X}\gamma\mathbf{d})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) + (\mathrm{X}\gamma\mathbf{d})'(\mathrm{X}\gamma\mathbf{d}) \\
&= -2(\mathrm{X}\mathbf{d})'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})\gamma + (\mathrm{X}\gamma\mathbf{d})'(\mathrm{X}\gamma\mathbf{d}) \\
&= -2\mathbf{d}'\mathrm{X}'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})\gamma + (\mathrm{X}\mathbf{d})'(\mathrm{X}\mathbf{d})\gamma^2 \\
&= -2\mathbf{d}'\mathrm{X}'(\mathbf{y} - \mathrm{X}\boldsymbol{\beta})\gamma + \mathbf{d}'\mathrm{X}'\mathrm{X}\mathbf{d}\gamma^2 \\
&= -2\mathbf{d}'\mathrm{c}\gamma + \mathbf{d}'\mathrm{X}'\mathrm{X}\mathbf{d}\gamma^2.
\end{aligned}
$$

□

Com as propriedades do algoritmo LARS expressas nos três lemas anteri-

ores e com uma sequência longa de lemas técnicos, é possível demonstrar que pelo algoritmo LARS se pode obter todas as estimativas LASSO, conforme em Efron et al. (2004, p. 11).

7 Estimação do Erro de Predição

Quando se ajusta a um conjunto de dados $\mathbf{y}$ (vetor de dimensão $\mathrm{n} \times 1$) um modelo $\mu(\mathbf{y})$ uma pergunta natural é o de se medir o quanto este modelo é adequado para predizer um novo conjunto de dados. Deseja-se então, ao menos, se ter um estimador para o erro de predição. A medida usual da soma dos quadrados dos resíduos do modelo $\|\mathbf{y} - \mu(\mathbf{y})\|^2$ é certamente otimista, uma vez que o modelo foi justamente ajustado com os dados observados $\mathbf{y}$. Tal questão pode ser rigorosamente colocada no sentido em que o que se quer é: Os dados $\mathbf{y}$ são aleatórios, o que implica que $\mu(\mathbf{y})$ também o é. Novos dados $\mathbf{y}^0$ também são aleatórios. O quadrado do erro de predição $\left\|\mathbf{y}^0 - \mu(\mathbf{y})\right\|^2$ também é aleatório. Um critério para se abordar este erro é tomar a esperança do quadrado do erro de predição. Observa-se que este erro depende de duas variáveis aleatórias independentes e igualmente distribuídas, e é dado pela esperança dupla $\mathrm{E}_0\left[\mathrm{E}\left[\left\|\mathbf{y}^0 - \mu(\mathbf{y})\right\|^2\right]\right]$.

7.1 Erro de predição para modelos lineares

O caso mais simples ocorre quando o modelo é linear e o estimador é não viesado, $\mu(\mathbf{y}) = \mathrm{M}\mathbf{y}$, com $\boldsymbol{\mu} = \mathrm{E}\left[\mathrm{M}\mathbf{y}\right] = \mathrm{M}\boldsymbol{\mu}$. Assim,

$$\begin{aligned} \mathrm{c} &= \mathrm{E}\left[\mathrm{E}_0\left[< \mathbf{y}^0 - \mathrm{M}\mathbf{y}, \mathbf{y}^0 - \mathrm{M}\mathbf{y} >\right]\right] \\ &= \mathrm{E}\left[\mathrm{E}_0\left[< \mathbf{y}^0, \mathbf{y}^0 > -2 < \mathbf{y}^0, \mathrm{M}\mathbf{y} > + < \mathrm{M}\mathbf{y}, \mathrm{M}\mathbf{y} >\right]\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\mathrm{E}_0\left[<\mathbf{y}^0,\mathbf{y}^0>\right]-2<\mathrm{E}_0\left[\mathbf{y}^0\right],\mathrm{M}\mathbf{y}>+<\mathrm{M}\mathbf{y},\mathrm{M}\mathbf{y}>\right]\\
&= \mathrm{E}\left[\|\boldsymbol{\mu}\|^2+\mathrm{n}\sigma^2-2<\boldsymbol{\mu},\mathrm{M}\mathbf{y}>+<\mathrm{M}\mathbf{y},\mathrm{M}\mathbf{y}>\right]\\
&= \|\boldsymbol{\mu}\|^2+\mathrm{n}\sigma^2-2<\boldsymbol{\mu},\mathrm{E}\left[\mathrm{M}\mathbf{y}\right]>+\mathrm{E}\left[<\mathrm{M}\mathbf{y},\mathrm{M}\mathbf{y}>\right]\\
&= \|\boldsymbol{\mu}\|^2+\mathrm{n}\sigma^2-2<\boldsymbol{\mu},\mathrm{M}\boldsymbol{\mu}>+\mathrm{E}\left[<\mathrm{M}\mathbf{y},\mathrm{M}\mathbf{y}>\right]\\
&= \|\boldsymbol{\mu}\|^2+\mathrm{n}\sigma^2-2<\boldsymbol{\mu},\mathrm{M}\boldsymbol{\mu}>+tr(\sigma^2\mathrm{M}'\mathrm{M})+\|\mathrm{M}\boldsymbol{\mu}\|^2.
\end{aligned}
$$

Mas como

$$
\begin{aligned}
\mathrm{E}\left[\|\mathbf{y}-\mathrm{M}\mathbf{y}\|^2\right] &= \mathrm{E}\left[<\mathbf{y}-\mathrm{M}\mathbf{y},\mathbf{y}-\mathrm{M}\mathbf{y}>\right]\\
&= \mathrm{E}\left[<\mathbf{y},\mathbf{y}>\right]-2\mathrm{E}\left[<\mathbf{y},\mathrm{M}\mathbf{y}>\right]+\mathrm{E}\left[<\mathrm{M}\mathbf{y},\mathrm{M}\mathbf{y}>\right]\\
&= \|\boldsymbol{\mu}\|^2+\mathrm{n}\sigma^2-2\mathrm{E}\left[<\mathbf{y},\mathrm{M}\mathbf{y}>\right]+\mathrm{E}\left[\mathbf{y}'\mathrm{M}'\mathrm{M}\mathbf{y}\right]\\
&= \|\boldsymbol{\mu}\|^2+\mathrm{n}\sigma^2-2\left\{\sigma^2 tr(\mathrm{M})+\boldsymbol{\mu}'\mathrm{M}\boldsymbol{\mu}\right\}\\
&\quad +\sigma^2 tr(\mathrm{M}'\mathrm{M})+\|\mathrm{M}\boldsymbol{\mu}\|^2,
\end{aligned}
$$

então $\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\mathrm{M}\mathbf{y}\right\|^2\right]\right] = \mathrm{E}\left[\|\mathbf{y}-\mathrm{M}\mathbf{y}\|^2\right]+2\sigma^2 tr\left(\mathrm{M}\right)$, de onde segue que um estimador não viesado para a esperança do quadrado do erro do modelo $\hat{\mu}=\mathrm{M}\mathbf{y}$ é dado por $\|\mathbf{y}-\mathrm{M}\mathbf{y}\|^2+2\sigma^2 tr(\mathrm{M})$. Uma vez que se tem a soma de resíduos observada do modelo e uma penalização dada por $2\sigma tr(\mathrm{M})^2$, tal penalização admite uma interpretação em termos de covariância, uma vez que

$$
\begin{aligned}
\mathrm{cov}\left(\mathbf{y},\mathrm{M}\mathbf{y}\right) &= \mathrm{E}\left[\left(\mathbf{y}-\boldsymbol{\mu}\right)'\left(\mathrm{M}\mathbf{y}-\boldsymbol{\mu}\right)\right]\\
&= \mathrm{E}\left[\left(\mathbf{y}-\boldsymbol{\mu}\right)'\left(\mathrm{M}\mathbf{y}\right)\right]\\
&= \mathrm{E}\left[\left(\mathbf{y}\right)'\left(\mathrm{M}\mathbf{y}\right)\right]-\mathrm{E}\left[\boldsymbol{\mu}'\mathrm{M}\mathbf{y}\right]\\
&= \boldsymbol{\mu}'\mathrm{M}\boldsymbol{\mu}+\sigma^2 tr(\mathrm{M})-\boldsymbol{\mu}'\mathrm{M}\boldsymbol{\mu}\\
&= \sigma^2 tr(\mathrm{M}).
\end{aligned}
$$

Desta forma um estimador não viesado para o erro de predição é dado por $\|\mathbf{y} - \mathrm{M}\mathbf{y}\|^2 + 2\mathrm{cov}\,(\mathbf{y}, \mathrm{M}\mathbf{y})$. Claramente este estimador não é operacional pois depende de um parâmetro populacional.

Em geral, os modelos lineares são dados por matrizes de projeção em um subespaço de dimensão p, que é justamente a dimensão do modelo. Neste caso, $tr(\mathrm{M}) = \mathrm{p}$ e o estimador fica da forma $\|\mathbf{y} - \mathrm{M}\mathbf{y}\|^2 + 2\mathrm{p}\sigma^2$. O estimador da esperança do quadrado do erro de predição evidentemente pode ser usado como uma ferramenta de seleção de modelos. De fato, o estimador não viesado obtido acima está intimamente relacionado à famosa fórmula do C_p de Mallows para seleção de modelos. Segue então uma demonstração de tal fórmula utilizando argumentos geométricos.

7.2 Estimação do erro de predição para o caso geral

No clássico artigo de Efron (2004), é apresentado um estimador não viesado para o erro de predição de modelos de regressão.

Considere $\mathbf{y}$ um vetor aleatório cujas coordenadas sejam as observações de um dado experimento. Se $\boldsymbol{\mu} = \mathrm{E}\,[\mathbf{y}]$, seja $\boldsymbol{\mu}(\mathbf{y})$ um estimador para a média $\boldsymbol{\mu}$ do vetor $\mathbf{y}$. Uma questão fundamental é mensurar a qualidade do estimador $\boldsymbol{\mu}(\mathbf{y})$ em medir futuras observações $\mathbf{y}^0$ geradas pelo mesmo mecanismo aleatório que gerou $\mathbf{y}$. Neste sentido, pode-se definir o erro de predição como o valor esperado do seguinte procedimento:

Observa-se os dados $\mathbf{y} \Rightarrow$ Obtém-se a estimativa da média $\boldsymbol{\mu}(\mathbf{y}) \Rightarrow$ Novos dados $\mathbf{y}^0$ do mesmo fenômeno são observados $\Rightarrow$ Calcula-se o quadrado do desvio $\left\|\mathbf{y}^0 - \boldsymbol{\mu}(\mathbf{y})\right\|^2$.

Portanto, o erro de predição é o parâmetro populacional (uma vez que é dado por uma esperança) definido pela esperança dupla $\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0 - \boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\right]$,

em que $\mathrm{E}[\ ]$ é a esperança em relação ao vetor aleatório $\mathbf{y}$ e $\mathrm{E}_0[\ ]$ representa a esperança em relação ao novo vetor aleatório $\mathbf{y}^0$. Um estimador não viesado para o erro de predição será obtido a partir da seguinte construção. Considere os triângulos conforme Figura 52.

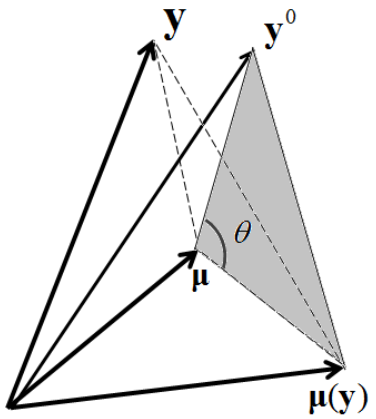


Figura 52 Geometria do erro de predição

Na Figura 52, é importante destacar novamente que o vetor $\mathbf{y}$ é o vetor cujas coordenadas são os valores observados, o vetor $\mathbf{y}^0$ é um vetor cujas coordenadas representam os valores de um novo experimento realizado e o vetor $\boldsymbol{\mu}$ é o vetor média populacional. Finalmente, $\boldsymbol{\mu}(\mathbf{y})$ é o vetor ajustado ao modelo, ou seja, é a projeção de $\mathbf{y}$ no modelo.

Aplicando-se a lei dos cossenos ao triângulo hachurado da Figura 52 tem-se:

$$\begin{aligned}\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2 &= \left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2+\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2 \\ &\quad -2\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|\cos(\theta) \\ &= \left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2+\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2-2<\mathbf{y}^0-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\end{aligned}$$

Tomando-se a esperança da equação em relação a $\mathbf{y}^0$

$$\begin{aligned} \mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right] &= \mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2+\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right. \\ &\quad \left.-2<\mathbf{y}^0-\mu,\mu(\mathbf{y})-\mu>\right] \\ &= \mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]+\mathrm{E}_0\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right] \\ &\quad -2\mathrm{E}_0\left[<\mathbf{y}^0-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\right] \end{aligned} \tag{7.1}$$

como $\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}$ não depende de $\mathbf{y}^0$ e uma vez que a esperança tomada é em relação à $\mathbf{y}^0$, pode-se escrever

$$\begin{aligned} -2\mathrm{E}_0\left[<\mathbf{y}^0-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\right] &= -2<\mathrm{E}_0\left[\mathbf{y}^0-\boldsymbol{\mu}\right],\mathrm{E}_0\left[\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right]> \\ &= -2<\mathrm{E}_0\left[\mathbf{y}^0-\boldsymbol{\mu}\right],\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}_0> \end{aligned}$$

e sendo $\mathrm{E}_0\left[\mathbf{y}^0-\boldsymbol{\mu}\right]=0$, então

$$-2\mathrm{E}_0\left[<\mathbf{y}^0-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\right]=-2<0,\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}_0>=0.$$

Portanto, a Equação 7.1 se escreve

$$\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]=\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]+\mathrm{E}_0\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right].$$

É interessante observar que, em termos de esperança vale o teorema de Pitágoras, isto é, em média o triângulo hachurado é um triângulo retângulo. Novamente, $\mathrm{E}_0\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right]=\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2$, pois $\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}$ não depende de $\mathbf{y}^0$. Logo, $\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]=\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]+\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2$, e portanto o erro de predição é dado por

$$\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\right] = \mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]\right]+\mathrm{E}\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right]. \tag{7.2}$$

$\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]$ é uma grandeza populacional que não depende de $\mathbf{y}$, e portanto $\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]\right]=\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]$. Logo, a Equação 7.2 fica

$$\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\right]=\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]+\mathrm{E}\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right].$$

Como $\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}\right\|^2\right]=\mathrm{E}\left[\left\|\mathbf{y}-\boldsymbol{\mu}\right\|^2\right]$ uma vez que $\mathbf{y}^0$ e $\mathbf{y}$ podem ser considerados a mesma variável aleatória, a Equação 7.2 se escreve

$$\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\right] \;=\; \mathrm{E}\left[\left\|\mathbf{y}-\boldsymbol{\mu}\right\|^2\right]+\mathrm{E}\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right] \tag{7.3}$$

Ao se aplicar a lei dos cossenos no triângulo tracejado da Figura 52 se tem

$$\begin{aligned}\mathrm{E}\left[\left\|\mathbf{y}-\boldsymbol{\mu}\right\|^2\right]+\mathrm{E}\left[\left\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\right\|^2\right] &= \mathrm{E}\left[\left\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\\ &\quad+2\mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\right],\end{aligned}$$

e portanto, substituindo na Equação 7.3

$$\begin{aligned}\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\right] &= \mathrm{E}\left[\left\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\\ &\quad+2\mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\right].\end{aligned}$$

Note que

$$\begin{aligned}\mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}>\right] &= \mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]+\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]-\boldsymbol{\mu}>\right]\\ &= \mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]>\right]\\ &\quad+\mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]-\boldsymbol{\mu}>\right]\\ &= \mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]>\right]\\ &\quad+\left[<\mathrm{E}\left[\mathbf{y}-\boldsymbol{\mu}\right],\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]-\boldsymbol{\mu}>\right]\\ &= \mathrm{E}\left[<\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\mu}(\mathbf{y})-\mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right]>\right]\end{aligned}$$

$$
\begin{aligned}
& + [< 0, \mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right] - \boldsymbol{\mu} >] \\
= \;& \mathrm{E}\left[< \mathbf{y} - \boldsymbol{\mu}, \boldsymbol{\mu}(\mathbf{y}) - \mathrm{E}\left[\boldsymbol{\mu}(\mathbf{y})\right] >\right] \\
= \;& \mathrm{E}\left[\sum_{i=1}^{n} \left(\mathbf{y}_i - \boldsymbol{\mu}_i\right) \left(\boldsymbol{\mu}(\mathbf{y})_i - \mathrm{E}[\boldsymbol{\mu}(\mathbf{y})]_i\right)\right] \\
= \;& \sum_{i=1}^{n} \mathrm{E}\left[\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right) \left(\boldsymbol{\mu}(\mathbf{y})_i - \mathrm{E}[\boldsymbol{\mu}(\mathbf{y})]_i\right)\right] \\
= \;& \sum_{i=1}^{n} \mathrm{E}\left[\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right) \left(\boldsymbol{\mu}(\mathbf{y})_i - \boldsymbol{\mu}_i\right)\right] \\
= \;& \sum_{i=1}^{n} \mathrm{cov}\left(\mathbf{y}_i, \boldsymbol{\mu}(\mathbf{y})_i\right) \\
= \;& \mathrm{cov}(\mathbf{y}, \boldsymbol{\mu}(\mathbf{y})),
\end{aligned}
$$

e então

$$
\mathrm{E}\left[\mathrm{E}_0\left[\left\|\mathbf{y}^0 - \boldsymbol{\mu}(\mathbf{y})\right\|^2\right]\right] = \mathrm{E}\left[\left\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{y})\right\|^2\right] + 2\mathrm{cov}(\mathbf{y}, \boldsymbol{\mu}(\mathbf{y})) \quad (7.4)
$$

O erro de predição é então a esperança do erro do ajuste do modelo acrescido de uma penalidade relativa à covariância entre os dados e o estimador.

Como exemplo, suponha que se quer calcular o erro de predição de um novo valor na seguinte situação: Seja $\mathbf{y} = (y_1, \ldots, y_n)$ uma amostra iid e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. O erro de predição para uma nova observação $\mathbf{y}^0$ é dado pela esperança dupla

$$
\begin{aligned}
E_0\left[E\left[\left\|\mathbf{y}^0 - \bar{\mathbf{y}}_n\right\|^2\right]\right] &= E_0\left[E\left[\left\|\mathbf{y}^0 - E_0[\mathbf{y}^0] + E_0[\mathbf{y}^0] - \bar{\mathbf{y}}_n\right\|^2\right]\right] \\
&= E_0 E\left[\left\|\mathbf{y}^0 - E_0[\mathbf{y}^0]\right\|^2\right] + E_0 E\left[\left\|E_0[\mathbf{y}^0] - \bar{\mathbf{y}}_n\right\|^2\right] \\
&\quad +2\left(\mathbf{y}^0 - E_0\left[\mathbf{y}^0\right]\right)\left(E_0\left[\mathbf{y}^0\right] - \bar{\mathbf{y}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= E_0 E\left[\left\|\mathbf{y}^0 - E_0[\mathbf{y}^0]\right\|^2\right] + E_0 E\left[\left\|E_0[\mathbf{y}^0] - \bar{\mathbf{y}}_n\right\|^2\right] \\
&\quad +2\left(\mathbf{y}^0 - E_0\left[\mathbf{y}^0\right]\right)(0) \\
&= E_0 E\left[\left\|\mathbf{y}^0 - E_0[\mathbf{y}^0]\right\|^2\right] + E E_0\left[\left\|E_0[\mathbf{y}^0] - \bar{\mathbf{y}}_n\right\|^2\right] \\
&= E_0 E\left[\left\|\mathbf{y}^0 - E_0[\mathbf{y}^0]\right\|^2\right] + E\left[\left\|E_0[\mathbf{y}^0] - \bar{\mathbf{y}}_n\right\|^2\right] \\
&= E_0\left[\left\|\mathbf{y}^0 - E_0[\mathbf{y}^0]\right\|^2\right] + E\left[\left\|E_0[\mathbf{y}^0] - \bar{\mathbf{y}}_n\right\|^2\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} \\
&= \frac{n+1}{n}\sigma^2. \qquad (7.5)
\end{aligned}
$$

Um outro exemplo é o erro de predição do estimador de quadrados mínimos em regressão linear. Chamando $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ de H, tem-se

$$
\begin{aligned}
E\left[E_0\left[\left\|\mathbf{y}_0 - \mathrm{H}\mathbf{y}\right\|^2\right]\right] &= E\left[E_0\left[< \mathbf{y}_0 - \mathrm{H}\mathbf{y}, \mathbf{y}_0 - \mathrm{H}\mathbf{y} >\right]\right] \\
&= E\left[E_0\left[< \mathbf{y}_0, \mathbf{y}_0 > -2 < \mathbf{y}_0, \mathrm{H}\mathbf{y} >\right.\right. \\
&\quad \left.\left. + < \mathrm{H}\mathbf{y}, \mathrm{H}\mathbf{y} >\right]\right] \\
&= E\left[E_0\left[< \mathbf{y}_0, \mathbf{y}_0 >\right] - 2E_0\left[< \mathbf{y}_0, \mathrm{H}\mathbf{y} >\right]\right. \\
&\quad \left. + E_0\left[< \mathrm{H}\mathbf{y}, \mathrm{H}\mathbf{y} >\right]\right] \\
&= E\left[E_0\left[< \mathbf{y}_0, \mathbf{y}_0 >\right]\right] - 2E\left[< E_0\left[\mathbf{y}_0\right], \mathrm{H}\mathbf{y} >\right] \\
&\quad + E\left[\left[< \mathrm{H}\mathbf{y}, \mathrm{H}\mathbf{y} >\right]\right] \\
&= E\left[E_0\left[< \mathbf{y}_0, \mathbf{y}_0 >\right]\right] - 2\left[< E_0\left[\mathbf{y}_0\right], \mathrm{H}E\left[\mathbf{y}\right] >\right] \\
&\quad + E\left[\left[< \mathrm{H}\mathbf{y}, \mathrm{H}\mathbf{y} >\right]\right] \\
&= E\left[E_0\left[< \mathbf{y}_0, \mathbf{y}_0 >\right]\right] - 2\left[< \boldsymbol{\mu}, \mathrm{H}\boldsymbol{\mu} >\right] \\
&\quad + E\left[\left[< \mathrm{H}\mathbf{y}, \mathrm{H}\mathbf{y} >\right]\right] \\
&= E_0\left[\mathbf{y}'_0 \mathbf{I} \mathbf{y}_0\right] - 2 < \boldsymbol{\mu}, \boldsymbol{\mu} > + E\left[\mathbf{y}' \mathrm{H} \mathbf{y}\right] \\
&= tr(\mathrm{I}\sigma^2\mathrm{I}) + \boldsymbol{\mu}'\mathrm{I}\boldsymbol{\mu} - 2\boldsymbol{\mu}'\mathrm{I}\boldsymbol{\mu} + tr(\mathrm{H}\sigma^2\mathrm{I}) + \boldsymbol{\mu}'\mathrm{H}\boldsymbol{\mu}
\end{aligned}
$$

$$
\begin{aligned}
&= n\sigma^2 - \boldsymbol{\mu'}\boldsymbol{\mu} + \sigma^2 tr(\mathrm{H}) + \boldsymbol{\mu'}\boldsymbol{\mu} \\
&= n\sigma^2 + \sigma^2 p \\
&= \sigma^2(n+p).
\end{aligned}
$$

Para se obter um estimador não viesado para o erro de predição, toma-se como estimador de $\mathrm{E}\left[\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{y})\|^2\right]$, utilizando o método dos momentos, a soma de quadrados dos resíduos $\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{y})\|^2$. É necessário agora se obter um estimador não viesado para $\mathrm{cov}(\mathbf{y}, \boldsymbol{\mu}(\mathbf{y}))$. Para tanto, utiliza-se o famoso Lema de Stein (1981 apud GAJO, 2016, p. 62).

Lema 1. (Stein) *Seja* $y \sim \mathrm{N}(\theta, 1)$. *Então*

$$
\mathrm{cov}\left(\mathrm{y}, \mathrm{h(y)}\right) = \mathrm{E}\left[\mathrm{h(y)(y} - \theta)\right] = \mathrm{E}\left[\mathrm{h'(y)}\right].
$$

Demonstração. Observe que

$$
\begin{aligned}
\mathrm{cov}\left(\mathrm{y}, \mathrm{h(y)}\right) &= \mathrm{E}\left[(y-\theta)\left(\mathrm{h}(y) - \mathrm{E}\left[\mathrm{h}(y)\right]\right)\right] \\
&= \mathrm{E}\left[(y-\theta)\mathrm{h}(y)\right] - \mathrm{E}\left[(y-\theta)\mathrm{E}\left[\mathrm{h}(y)\right]\right] \\
&= \mathrm{E}\left[(y-\theta)\mathrm{h}(y)\right] - \mathrm{E}\left[(y-\theta)\right]\mathrm{E}\left[\mathrm{h}(y)\right] \\
&= \mathrm{E}\left[(y-\theta)\mathrm{h}(y)\right].
\end{aligned}
$$

Utilizando integração por partes

$$
\begin{aligned}
\mathrm{E}\left[\mathrm{h}(y)(y-\theta)\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{h}(y)(y-\theta) \exp\left[-\frac{1}{2}(y-\theta)^2\right] dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{h}(y) \left[\frac{d}{dy}\left(-\exp\left[-\frac{1}{2}(y-\theta)^2\right]\right)\right] dy \\
&= -\frac{1}{\sqrt{2\pi}} \mathrm{h}(y) \exp\left[-\frac{1}{2}(y-\theta)^2\right] \Bigg|_{-\infty}^{\infty}
\end{aligned}
$$

$$
\begin{aligned}
& + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{h}'(y) \exp\left[-\frac{1}{2}(y-\theta)^2\right] dy \\
= \; & \mathrm{E}\left[\mathrm{h}'(y)\right].
\end{aligned}
$$

□

O lema pode ser facilmente estendido para variável aleatória normal e variância qualquer. Se $\mathrm{y} \sim \mathrm{N}\left(\theta, \sigma^2\right)$, seja $x = \frac{\mathrm{y}-\theta}{\sigma} \sim \mathrm{N}\left(0,1\right)$. Como $\mathrm{y} = \sigma x + \theta$, então

$$\mathrm{h(y)} = \mathrm{h}\left(\sigma x + \theta\right) = g(x)$$

e,

$$
\begin{aligned}
\mathrm{E}\left[\mathrm{h}(y)\left(\frac{y-\theta}{\sigma}\right)\right] &= \mathrm{E}\left[g(x)x\right] \\
&= \mathrm{E}\left[g'(x)\right] \quad \text{(pelo lema)} \\
&= \mathrm{E}\left[\sigma h'(\sigma x + \theta)\right] \\
&= \mathrm{E}\left[\sigma h'(y)\right].
\end{aligned}
$$

Assim, $\frac{1}{\sigma^2}\mathrm{E}\left[\mathrm{h}(y)\left(y-\theta\right)\right] = \mathrm{E}\left[\mathrm{h}'(\mathrm{y})\right]$.

O Lema de Stein admite uma generalização para variáveis normais multivariadas $\mathbf{y} \sim \mathrm{N_n}(\boldsymbol{\mu}, \sigma^2\mathrm{I})$. No entanto, para esta generalização é necessário supor que a função $\mathrm{h}:\mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}$ admita uma propriedade técnica denominada *quasi diferenciável*. As funções que têm esta propriedade possuem derivadas parciais em quase todo ponto. Denominando $\nabla h = \left(\frac{\partial h}{\partial y_1}, \ldots, \frac{\partial h}{\partial y_{\mathrm{n}}}\right)$ tem-se o Lema de Stein Multivariado (STEIN, 1981 apud GAJO, 2016, p. 66).

Lema 2. *Seja* $\mathbf{y} \sim \mathrm{N_n}(\mu, \sigma^2\mathrm{I})$ *e* $\mathrm{h}:\mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}$ *quasi diferenciável. Então*

$$\frac{1}{\sigma^2}\mathrm{E}\left[\mathrm{h}(\mathbf{y})\left(\mathbf{y}-\boldsymbol{\mu}\right)\right] = \mathrm{E}\left[\nabla \mathrm{h}(\mathbf{y})\right].$$

Como observação final, o lema de Stein pode ser ainda colocado em termos mais gerais. Considere novamente $\mathbf{y} \sim \mathrm{N_n}\left(\mu, \sigma^2 \mathrm{I}\right)$ e $f : \mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}^{\mathrm{n}}$, $f = (f_1, \ldots f_n)$ um campo vetorial. Pelo Lema 2, então

$$\frac{1}{\sigma^2}\mathrm{E}\left[(\mathbf{y} - \mu)\, f_i(\mathbf{y})\right] = \mathrm{E}\left[\nabla f_i(\mathbf{y})\right].$$

Portanto, $\frac{1}{\sigma^2}\mathrm{cov}\left(\mathrm{y}_i, f_i(\mathbf{y})\right) = \frac{1}{\sigma^2}\mathrm{E}\left[(\mathrm{y}_i - \mu_i)\, f_i(\mathbf{y})\right] = \mathrm{E}\left[\frac{\partial}{\partial y_i} f_i(\mathbf{y})\right]$. Somando em i tem-se

$$\begin{aligned}
\frac{1}{\sigma^2}\mathrm{cov}(\mathbf{y}, f(\mathbf{y})) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{cov}(y_i, f_i(\mathbf{y})) \\
&= \mathrm{E}\left[\sum_{i=1}^{n}\frac{\partial}{\partial y_i} f_i(\mathbf{y})\right] \\
&= \mathrm{E}\left[div\, f(\mathbf{y})\right].
\end{aligned}$$

Como consequência do Lema de Stein, tem-se um estimador não viesado para $\mathrm{cov}\left(\mathbf{y}, \mu(\mathbf{y})\right)$ dado por $div\, \mu(\mathbf{y}) = \sum_{i=1}^{n}\frac{\partial}{\partial y_i}\mu_i(\mathbf{y})$. Utilizando este estimador obtém-se a partir da Equação 7.4 um estimador não viesado para o erro de predição

$$\|\mathbf{y} - \mu(\mathbf{y})\|^2 + 2\sigma^2 div\, \mu(\mathbf{y}) = \|\mathbf{y} - \mu(\mathbf{y})\|^2 + 2\sigma^2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\mu_i(\mathbf{y}).$$

Definição 3. *O Grau de Liberdade, ou degree of freedom (df), de um modelo $\mu(\mathbf{y})$ é a grandeza populacional*

$$\mathrm{df} = \frac{\mathrm{cov}\left(\mathbf{y}, \mu(\mathbf{y})\right)}{\sigma^2} = \sum_{i=1}^{n}\frac{\mathrm{cov}\left(\mathrm{y}_i, \mu_i(\mathbf{y})\right)}{\sigma^2}.$$

Portanto, um estimador não viesado para o grau de liberdade é $\widehat{df} = \sum_{i=1}^{n}\frac{\partial}{\partial y_i}\mu_i(\mathbf{y})$. Uma outra importante aplicação do Lema de Stein é a obtenção de um estimador não viesado para o risco de estimadores. O risco de um estimador é a esperança da função perda. No caso da função perda quadrática, o risco é nada

mais do que o erro quadrático médio. Aplicando a lei dos cossenos ao triângulo tracejado da Figura 52,

$$\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{y})\|^2 = \|\mathbf{y}-\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\|^2 - 2 < \mathbf{y}-\boldsymbol{\mu}, \boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu} >$$

$$\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\|^2 = \|\mathbf{y}-\boldsymbol{\mu}(\mathbf{y})\|^2 - \|\mathbf{y}-\boldsymbol{\mu}\|^2 + 2 < \mathbf{y}-\boldsymbol{\mu}, \boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu} >$$

Logo,

$$\begin{aligned}
\Re &= \mathrm{E}\left[\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\|^2\right] \\
&= \mathrm{E}\left[\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{y})\|^2\right] - \mathrm{E}\left[\|\mathbf{y}-\boldsymbol{\mu}\|^2\right] + 2\mathrm{E} < \mathbf{y}-\boldsymbol{\mu}, \boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu} > \\
&= \|\mathbf{y}-\boldsymbol{\mu}(\mathbf{y})\|^2 - \mathrm{n}\sigma^2 + 2\mathrm{cov}(\mathbf{y}, \boldsymbol{\mu}(\mathbf{y})).
\end{aligned}$$

E portanto, um estimador não viesado para o risco, isto é, para o erro de predição, é

$$\hat{\Re} = -n\sigma^2 + \|\boldsymbol{\mu}(\mathbf{y})-\mathbf{y}\|^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial}{\partial y_i}\mu_i(\mathbf{y}).$$

Este estimador é denominado na literatura de SURE (*Stein's Unbiased Risk Estimate*).

Exemplo 1: Considere $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_n)$ uma amostra iid com $\mathrm{var}(\mathrm{y}_i) = \sigma^2$ e $\boldsymbol{\mu}(\mathbf{y}) = (\bar{\mathrm{y}}, \ldots, \bar{\mathrm{y}})$ com $\bar{\mathrm{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{y}_i$. Logo,

$$\begin{aligned}
\hat{\Re} &= -n\sigma^2 + \|\boldsymbol{\mu}(\mathbf{y})-\mathbf{y}\|^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial}{\partial y_i}\bar{\mathrm{y}} \\
&= -n\sigma^2 + \sum_{i=1}^{n} (\mathrm{y}_i - \bar{\mathrm{y}})^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial}{\partial y_i}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{y}_i\right) \\
&= -n\sigma^2 + \sum_{i=1}^{n} (\mathrm{y}_i - \bar{\mathrm{y}})^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{1}{n} \\
&= -n\sigma^2 + \sum_{i=1}^{n} (\mathrm{y}_i - \bar{\mathrm{y}})^2 + 2\sigma^2
\end{aligned}$$

Observe que $\mathrm{E}\left[\sum_{i=1}^{n}(\mathrm{y}_i-\bar{\mathrm{y}})^2\right]=(n-1)\sigma^2$. Logo $\mathrm{E}\left[\hat{\Re}\right]=\sigma^2$.

Exemplo 2: Considere o problema de regressão linear múltipla $\mathbf{y}=\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ em que X é uma matriz $\mathrm{n}\times\mathrm{p}$ de posto completo, $\mathbf{y}=(\mathrm{y}_1,\ldots,\mathrm{y}_n)$ e $\mathrm{var}(\mathrm{y}_i)=\sigma^2$. Considerando o subespaço $\mathrm{Im}(\mathrm{X})$ como o modelo, o estimador $\mu(\mathbf{y})$ é dado pela projeção ortogonal de $\mathbf{y}$ em $\mathrm{Im}(\mathrm{X})$. Esta projeção é dada explicitamente pela matriz de projeção (*Hat Matrix*) $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ e $\mu(\mathbf{y})=\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$. Logo,

$$
\begin{aligned}
\hat{\Re} &= -n\sigma^2+\|\boldsymbol{\mu}(\mathbf{y})-\mathbf{y}\|^2+2\sigma^2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\boldsymbol{\mu}(\mathbf{y})\\
&= -n\sigma^2+\left\|\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\mathbf{y}\right\|^2+2\sigma^2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\Big(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathbf{y})\Big)_i\\
&= -n\sigma^2+\left\|\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\mathbf{y}\right\|^2+2\sigma^2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\sum_{j=1}^{n}\left[\Big(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\Big)_{\mathrm{ij}}\mathrm{y_j}\right]\\
&= -n\sigma^2+\left\|\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\mathbf{y}\right\|^2+2\sigma^2\sum_{i=1}^{n}\Big(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\Big)_{\mathrm{ii}}\\
&= -n\sigma^2+\left\langle\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\mathbf{y},\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\mathbf{y}\right\rangle+2\sigma^2 tr(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')\\
&= -n\sigma^2+<\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y},\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}>\\
&\quad -2<\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y},\mathbf{y}>+<\mathbf{y},\mathbf{y}>+2\sigma^2 tr(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')\\
&= -n\sigma^2+\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-2\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}+\mathbf{y}'\mathbf{y}+2\mathrm{p}\sigma^2\\
&= -n\sigma^2+\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-2\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}+\mathbf{y}'\mathbf{y}+2\mathrm{p}\sigma^2\\
&= -n\sigma^2-\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}+\mathbf{y}'\mathbf{y}+2\mathrm{p}\sigma^2\\
&= (2\mathrm{p}-\mathrm{n})\sigma^2+\mathbf{y}'\left(\mathrm{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathbf{y}.
\end{aligned}
$$

De fato,

$$
\begin{aligned}
\Re &= \mathrm{E}\left[\|\boldsymbol{\mu}(\mathbf{y})-\boldsymbol{\mu}\|^2\right]\\
&= \mathrm{E}\left[\left\langle\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\boldsymbol{\mu},\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}-\boldsymbol{\mu}\right\rangle\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{E}\left[\left\langle \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}, \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right\rangle - 2\left\langle \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}, \boldsymbol{\mu}\right\rangle + \langle\boldsymbol{\mu},\boldsymbol{\mu}\rangle\right] \\
&= \mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - 2\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu}\right] \\
&= \mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} - 2\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu}\right] \\
&= \mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right] - 2\mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu}\right] + \boldsymbol{\mu}'\boldsymbol{\mu} \\
&= \mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right] - 2\mathrm{E}\left[\mathbf{y}'\right]\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu} \\
&= \mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right] - 2\boldsymbol{\mu}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu} \\
&= \mathrm{E}\left[\mathbf{y}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right] - 2\boldsymbol{\mu}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu} \\
&= tr(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\sigma^2\mathrm{I}) + \boldsymbol{\mu}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\mu} - \boldsymbol{\mu}'\boldsymbol{\mu} \quad (\mathrm{R}encher 5.2.a) \\
&= \mathrm{p}\sigma^2 + \boldsymbol{\mu}'\boldsymbol{\mu} - \boldsymbol{\mu}'\boldsymbol{\mu} = \mathrm{p}\sigma^2.
\end{aligned}
$$

Exemplo 3: O resultado fundamental de que o estimador de James-Stein para a média de uma normal multivariada domina estritamente a média amostral pode ser facilmente demonstrado utilizando o SURE. Para $\mathbf{y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \mathrm{I})$ o estimador de James-Stein tem a expressão $\mu_{\mathrm{JS}}(\mathbf{y}) = \left(1 - \frac{n-2}{\|\mathbf{y}\|^2}\right)\mathbf{y}$. O SURE para este estimador é

$$
\begin{aligned}
\hat{\Re} &= -n\sigma^2 + \|\boldsymbol{\mu}(\mathbf{y}) - \mathbf{y}\|^2 + 2\sigma^2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\boldsymbol{\mu}(\mathbf{y}) \\
&= -n + \left\|\left(1 - \frac{n-2}{\|\mathbf{y}\|^2}\right)\mathbf{y} - \mathbf{y}\right\|^2 + 2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\left[\left(1 - \frac{n-2}{\|\mathbf{y}\|^2}\right)\mathbf{y}\right] \\
&= -n + \left\|\left(\frac{n-2}{\|\mathbf{y}\|^2}\right)\mathbf{y}\right\|^2 + 2\sum_{i=1}^{n}\frac{\partial}{\partial y_i}\left[\left(1 - \frac{n-2}{\|\mathbf{y}\|^2}\right)\mathbf{y}\right] \\
&= -n + \left(\frac{n-2}{\|\mathbf{y}\|^2}\right)^2\|\mathbf{y}\|^2 + 2\sum_{i=1}^{n}\left[1 - \frac{n-2}{\|\mathbf{y}\|^2} + \frac{n-2}{\|\mathbf{y}\|^4}2(\mathrm{y}_i)^2\right] \\
&= -n + \frac{(n-2)^2}{\|\mathbf{y}\|^2} + 2\left(n - n\frac{n-2}{\|\mathbf{y}\|^2} + \frac{n-2}{\|\mathbf{y}\|^4}2\sum_{i=1}^{n}(\mathrm{y}_i)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -n + \frac{(n-2)^2}{\|\mathbf{y}\|^2} + 2\left(n - n\frac{n-2}{\|\mathbf{y}\|^2} + \frac{n-2}{\|\mathbf{y}\|^4}2\|\mathbf{y}\|^2\right) \\
&= -n + \frac{(n-2)^2}{\|\mathbf{y}\|^2} + 2\left(n - \frac{(n-2)}{\|\mathbf{y}\|^2}(n-2)\right) \\
&= -n + \frac{(n-2)^2}{\|\mathbf{y}\|^2} + 2\left(n - \frac{(n-2)^2}{\|\mathbf{y}\|^2}\right) \\
&= n - \frac{(n-2)^2}{\|\mathbf{y}\|^2}.
\end{aligned}
$$

Portanto, o risco do estimador de James-Stein é

$$
\mathrm{E}\left[\hat{\Re}\right] = \mathrm{E}\left[n - \frac{(n-2)^2}{\|\mathbf{y}\|^2}\right] = n - (n-2)^2\mathrm{E}\left[\frac{1}{\|\mathbf{y}\|^2}\right] < n = \mathrm{E}\left[\|\mathbf{y} - \mu\|^2\right],
$$

ou seja, o risco do estimador de James-Stein é estritamente menor do que o risco do estimador natural, isto é, a própria observação $\mathbf{y}$.

7.3 Seleção de modelos

7.3.1 A Estatística C_p de Mallows

O uso da estatística C_p de Mallow's é uma técnica para seleção de modelos em regressão. Define um critério de ajuste para comparação de modelos com diferente número de parâmetros. Para modelo linear M, $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, com $\mathrm{E}[\mathbf{y}] = \mathrm{X}\boldsymbol{\beta} = \boldsymbol{\theta}$, se supõe que o parâmetro esperança de $\mathbf{y}$ pertença ao subespaço linear $\mathrm{Im(X)}$. Se o modelo não for correto, então $\mathrm{E}[\mathbf{y}] = \boldsymbol{\theta} \notin \mathrm{Im(X)}$ e pode-se decompor ortogonalmente $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{X}} + \boldsymbol{\theta}_{\mathrm{X}^\perp}$ em relação ao subespaço $\mathrm{Im(X)}$. A estimação pelo modelo errado é dada pela projeção ortogonal em $\mathrm{Im(X)}$, $\mathrm{P}_{\mathrm{Im(X)}}\mathbf{y} = \boldsymbol{\theta}_{\mathrm{X}} + \boldsymbol{\varepsilon}_{\mathrm{X}}$. Desta forma, a matriz de dispersão deste estimador é

$$
\mathrm{E}\left[(\hat{\mathbf{y}} - \boldsymbol{\theta})(\hat{\mathbf{y}} - \boldsymbol{\theta})'\right] = \mathrm{E}\left[(\boldsymbol{\theta}_{\mathrm{X}} + \boldsymbol{\varepsilon}_{\mathrm{X}} - (\boldsymbol{\theta}_{\mathrm{X}} + \boldsymbol{\theta}_{\mathrm{X}^\perp}))\right.
$$

$$
\begin{aligned}
& (\boldsymbol{\theta}_{\mathrm{X}} + \boldsymbol{\varepsilon}_{\mathrm{X}} - (\boldsymbol{\theta}_{\mathrm{X}} + \boldsymbol{\theta}_{\mathrm{X}^{\perp}}))' \big] \\
= \; & \mathrm{E}\left[(\boldsymbol{\varepsilon}_{\mathrm{X}} - \boldsymbol{\theta}_{\mathrm{X}^{\perp}}) (\boldsymbol{\varepsilon}_{\mathrm{X}} - \boldsymbol{\theta}_{\mathrm{X}^{\perp}})' \right] \\
= \; & \mathrm{E}\left[\boldsymbol{\varepsilon}_{\mathrm{X}} \boldsymbol{\varepsilon'}_{\mathrm{X}} + \boldsymbol{\theta}_{\mathrm{X}^{\perp}} \boldsymbol{\theta'}_{\mathrm{X}^{\perp}} \right] \\
= \; & \mathrm{E}\left[\boldsymbol{\varepsilon}_{\mathrm{X}} \boldsymbol{\varepsilon'}_{\mathrm{X}} \right] + \boldsymbol{\theta}_{\mathrm{X}^{\perp}} \boldsymbol{\theta'}_{\mathrm{X}^{\perp}} \\
= \; & \sigma^2 \mathbf{I}_{p \times p} + \boldsymbol{\theta}_{\mathrm{X}^{\perp}} \boldsymbol{\theta'}_{\mathrm{X}^{\perp}}.
\end{aligned}
$$

O erro quadrático médio do estimador $\hat{\mathbf{y}}$ é dado por

$$
\begin{aligned}
\mathrm{EQM}(\hat{\mathbf{y}}) &= \mathrm{E}\left[(\hat{\mathbf{y}} - \boldsymbol{\theta})' (\hat{\mathbf{y}} - \boldsymbol{\theta}) \right] \\
&= tr\left(\mathrm{E}\left[(\hat{\mathbf{y}} - \boldsymbol{\theta})' (\hat{\mathbf{y}} - \boldsymbol{\theta}) \right] \right) \\
&= tr\left(\mathrm{E}\left[(\hat{\mathbf{y}} - \boldsymbol{\theta}) (\hat{\mathbf{y}} - \boldsymbol{\theta})' \right] \right) \\
&= tr\left[\sigma^2 \mathbf{I} + \boldsymbol{\theta}_{\mathrm{X}^{\perp}} \boldsymbol{\theta'}_{\mathrm{X}^{\perp}} \right] \\
&= \mathrm{p}\sigma^2 + \| \boldsymbol{\theta}_{\mathrm{X}^{\perp}} \|^2.
\end{aligned}
$$

Ao se optar entre vários modelos, deve-se escolher aquele que tem menor $\mathrm{EQM}(\hat{\mathbf{y}})$. Entretanto, uma vez que $\mathrm{EQM}(\hat{\mathbf{y}})$ depende de parâmetros populacionais, é necessário buscar um estimador não viesado para esta grandeza. Primeiramente, calcula-se a esperança da estatística $\hat{\sigma}^2 = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\mathrm{n-p}}$, por

$$
\begin{aligned}
\mathrm{E}\left[\hat{\sigma}^2 \right] &= \mathrm{E}\left[\frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\mathrm{n-p}} \right] \\
&= \mathrm{E}\left[\frac{\mathbf{y}' (\mathrm{I} - \mathrm{P_X}) \mathbf{y}}{\mathrm{n-p}} \right] \\
&= \frac{1}{\mathrm{n-p}} \mathrm{E}\left[(\boldsymbol{\theta} + \boldsymbol{\varepsilon})' (\mathrm{I} - \mathrm{P_X}) (\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \right] \\
&= \frac{1}{\mathrm{n-p}} \mathrm{E}\left[tr\left\{ (\mathrm{I} - \mathrm{P_X}) (\boldsymbol{\theta} + \boldsymbol{\varepsilon}) (\boldsymbol{\theta} + \boldsymbol{\varepsilon})' \right\} \right] \\
&= \frac{1}{\mathrm{n-p}} \mathrm{E}\left[tr\left\{ (\boldsymbol{\theta}_{\mathrm{X}^{\perp}} + \boldsymbol{\varepsilon}_{\mathrm{X}^{\perp}}) (\boldsymbol{\theta} + \boldsymbol{\varepsilon})' \right\} \right]
\end{aligned}
$$

$$\begin{aligned} &= \frac{1}{\mathrm{n}-\mathrm{p}}\mathrm{E}\left[\left(\boldsymbol{\theta}_{\mathrm{X}^{\perp}}+\boldsymbol{\varepsilon}_{\mathrm{X}^{\perp}}\right)\left(\boldsymbol{\theta}_{\mathrm{X}}+\boldsymbol{\varepsilon}_{\mathrm{X}}\right)'\right] \\ &= \frac{1}{\mathrm{n}-\mathrm{p}}\left(\boldsymbol{\theta}_{\mathrm{X}^{\perp}}\boldsymbol{\theta'}_{\mathrm{X}^{\perp}}+\mathrm{E}\left[\boldsymbol{\varepsilon}_{\mathrm{X}^{\perp}}\boldsymbol{\varepsilon'}_{\mathrm{X}^{\perp}}\right]\right) \\ &= \frac{1}{\mathrm{n}-\mathrm{p}}\left\|\boldsymbol{\theta}_{\mathrm{X}^{\perp}}\right\|^2+\sigma^2. \end{aligned}$$

Logo, para $(\mathrm{n}-\mathrm{p})\,\hat{\sigma}^2-(\mathrm{n}-2\mathrm{p})\,\sigma^2$, tem-se

$$\begin{aligned} \mathrm{E}\left[(\mathrm{n}-\mathrm{p})\,\hat{\sigma}^2-(\mathrm{n}-2\mathrm{p})\,\sigma^2\right] &= \left\|\boldsymbol{\theta}_{\mathrm{X}}\right\|^2+(\mathrm{n}-\mathrm{p})\,\sigma^2-(\mathrm{n}-2\mathrm{p})\,\sigma^2 \\ &= \left\|\boldsymbol{\theta}_{\mathrm{X}}\right\|^2+\mathrm{p}\sigma^2 \\ &= \mathrm{EQM}(\hat{\mathbf{y}}). \end{aligned}$$

Portanto, $\hat{\sigma}^2$ é um estimador viesado de σ^2. Se é obtido um estimador não viesado $\hat{\sigma}^*$ para σ^2, por exemplo, considerando um modelo de dimensão muito maior do que o modelo M considerado, se tem então um estimador não viesado para $\mathrm{EQM}(\hat{\mathbf{y}})$ dado por

$$\begin{aligned} \mathrm{EQM}(\hat{\mathbf{y}}) &= \mathrm{E}\left[(\mathrm{n}-\mathrm{p})\left(\hat{\sigma}^2-\hat{\sigma}^{2*}\right)+\mathrm{p}\hat{\sigma}^2\right] \\ &= \mathrm{E}\left[\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2-(\mathrm{n}-\mathrm{p})\,\hat{\sigma}^{2*}+\mathrm{p}\sigma^2\right]. \end{aligned}$$

Desta forma, $\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2-(\mathrm{n}-\mathrm{p})\,\hat{\sigma}^{2*}+\mathrm{p}\hat{\sigma}^{2*}=\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2+2\mathrm{p}\hat{\sigma}^{2*}-\mathrm{n}\hat{\sigma}^{2*}$ é um estimador não viesado de $\mathrm{EQM}(\hat{\mathbf{y}})$. A estatística $\mathrm{C_p}$ de Mallows é então definida como

$$C_p=\frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2+2\mathrm{p}\hat{\sigma}^{2*}-\mathrm{n}\hat{\sigma}^{2*}}{\hat{\sigma}^{2*}}=\frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2}{\hat{\sigma}^{2*}}+2\mathrm{p}-\mathrm{n}=\frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2}{\hat{\sigma}^{2*}}-(\mathrm{n}-2\mathrm{p}).$$

Assim, a estatística C_p é um estimador não viesado de $\frac{1}{\mathrm{n}-\mathrm{p}}\left\|\theta_{\mathrm{X}^{\perp}}\right\|^2+\sigma^2$, isto é, de uma medida ponderada pela dimensão de quanto o modelo se afasta do modelo real. No caso de se escolher entre modelos, deve-se optar por aquele

com menor C_p, pois neste caso é razoável admitir que este possui o menor viés ponderado pela dimensão. A constituição deste estimador pode ser explicada pela geometria da construção (Figura 53).

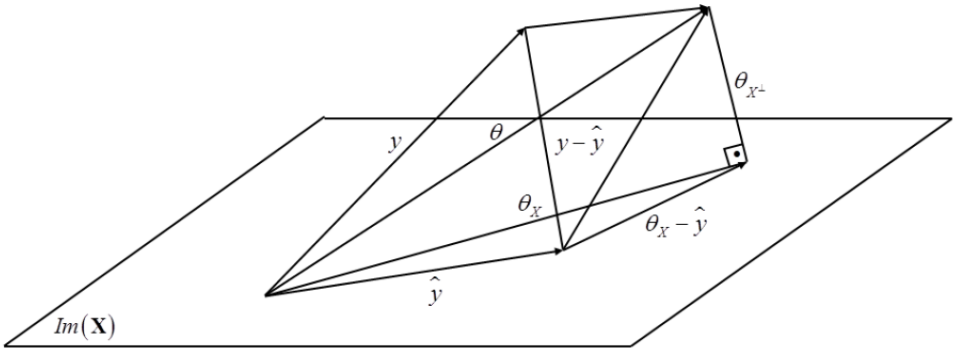


Figura 53 Elementos representativos da estrutura $\mathrm{C_p}$.

7.4 O conceito de grau de liberdade

Grau de liberdade é um dos termos mais utilizados na literatura estatística, sendo onipresente nos artigos científicos da área. No entanto, a experiência mostra que o conceito é utilizado sem que de fato se tenha consciência de seu verdadeiro significado. Uma razão para tal fato, talvez, seja que o grau de liberdade pode e é utilizado em vários sentidos. Vejamos alguns deles. Ao se referir a uma qui-quadrado com n graus de liberdade, ou a uma distribuição t com n graus de liberdade, n significa simplesmente um parâmetro para estas distribuições, ou seja, ambas as distribuições são famílias uniparamétricas indexadas pelos números naturais, os quais são denominados Grau de Liberdade. A origem do termo decorre do fato de que uma variável aleatória obtida como soma de quadrados de n normais padrão independentes tem distribuição qui-quadrado com n graus de liberdade. Nesse sentido, pode-se supor que, se é observado um fenômeno que depende, por exemplo, de n fatores e estes se combinem de forma que nenhum de-

les restrinja o comportamento dos demais, então é razoável dizer que este sistema tem n graus de liberdade. Como definir então grau de liberdade para métodos de estimação mais gerais?

No caso de regressão linear $\mathbf{y}_{n\times 1} = \mathrm{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, com X de posto completo e $p < n$, o grau de liberdade é a dimensão da imagem da transformação X, isto é, a dimensão do modelo. O grau de liberdade, em uma definição geral, deve refletir isto, a complexidade do modelo.

O modelo supõe que todos os valores das componentes do vetor de parâmetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ podem assumir qualquer valor real. Isto significa que a única suposição sobre o vetor de médias $\mu = (\mu_1, \ldots, \mu_n) = \mathrm{E}[\mathbf{y}]$ é que este pertença ao subespaço imagem de X (denotado por $\mathrm{Im}(\mathrm{X})$). A dimensão deste subespaço é p. O processo de estimação derivado do método dos quadrados mínimos é simplesmente a projeção ortogonal do vetor observado $\mathbf{y}$ no subespaço $\mathrm{Im}(\mathrm{X})$. Observe que todas as amostras da forma $\mathbf{y} + \mathbf{w}$, com $\mathbf{w}$ pertencendo ao subespaço ortogonal à $\mathrm{Im}(\mathrm{X})$, resultam na mesma estimativa para o vetor de médias, uma vez que o vetor $\mathbf{w}$ é projetado na origem. Desta forma, estas amostras não contêm novas informações para a estimação do vetor de médias. Como, informalmente, o número de tais amostras, isto é, o número de vetores $\mathbf{w}$, pode ser medido pela dimensão do espaço ortogonal, que é $n - p$, então a quantidade de informação da amostra em relação à estimação da média fica restrita à $\mathrm{Im}(\mathrm{X}) = p$. Logo, o grau de liberdade pode ser explicado também em termos da dimensão do modelo, isto é, a dimensão do espaço em que se supões estar o vetor de médias μ. No aspecto algébrico, um subespaço de dimensão k é definido por k equações lineares. Escolhendo-se uma base formada por vetores nesse subespaço e por vetores no ortogonal, o subespaço é simplesmente definido zerando-se as variáveis correspondentes ao subespaço ortogonal, ficando livres as demais variáveis. Esta

construção leva à definição mais utilizada de grau de liberdade, que é a diferença entre o número total de variáveis e o número de vínculos, ou seja, o número de equações. Portanto, a interpretação geométrica é clara no sentido de que o grau de liberdade é igual à dimensão de $\mathrm{Im(X)} = p$, que é o número de variáveis livres para determinação do vetor de médias μ, sendo que as variáveis relativas ao subespaço ortogonal de $\mathrm{Im(X)}$ são dependentes desta.

Como visto na seção relativa à estimação do erro de predição de um modelo, vê-se que a esperança do erro de predição é o erro quadrático médio mais uma penalização dada por $(\mathbf{y}, \mu(\mathbf{y}))$. Como também foi visto para o caso linear $\mu(\mathbf{y}) = \mathrm{My}$, $\frac{\mathrm{cov}(\mathbf{y},\mu(\mathbf{y}))}{\sigma^2}$ é exatamente $\mathrm{tr(M)}$ e, portanto, o grau de liberdade usual.

Certamente a penalização $\mathrm{cov}(\mathbf{y}, \mu(\mathbf{y}))$ reflete a complexidade do modelo. A ideia de complexidade e a de graus de liberdade em um sistema físico estão intuitivamente relacionadas. Pensando em um sistema físico a complexidade, e portanto o grau de liberdade, não deve depender da amplitude de variação do sistema. Um sistema planetário não é necessariamente mais complexo do que um sistema orbital de um átomo. Portanto, é razoável que $\mathrm{cov}(\mathbf{y}, \mu(\mathbf{y}))$ seja dividido por σ^2 na penalização do erro de predição. Com as considerações acima descritas é razoável então a Definição 3 (EFRON et al., 2004) na Subseção 7.2. A partir desta definição, df é evidentemente um parâmetro populacional e precisa ser estimado. Neste sentido tem-se o resultado fundamental dado pelo Lema 1 (Stein). Sob suposição de normalidade $\mathrm{N_n}(\mu, \sigma^2\mathrm{I})$, tem-se que

$$\begin{aligned}
\frac{\mathrm{cov}(\mathbf{y}, \mu(\mathbf{y}))}{\sigma^2} &= \frac{1}{\sigma^2}\mathrm{E}\left[(\mathbf{y} - \boldsymbol{\mu})\left(\mu(\mathbf{y}) - \mathrm{E}[\mu(\mathbf{y})]\right)\right] \\
&= \frac{1}{\sigma^2}\mathrm{E}\left[(\mathbf{y} - \boldsymbol{\mu})\mu(\mathbf{y})\right] \\
&= \mathrm{E}\left[\nabla\mu(\mathbf{y})\right],
\end{aligned}$$

e também $\nabla \mathrm{g}(\mathbf{x}) = \mathrm{div}\, \mathrm{g}(\mathbf{x}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial \mathrm{g_i}(\mathbf{x})}{\partial \mathrm{x_i}}$ é o divergente da função g.

Pelo lema de Stein tem-se então como estimador não viesado de df de um modelo $\mathrm{M}(\mathbf{y})$, $\nabla \mathrm{M}(\mathbf{y}) = \operatorname{div} \mu(\mathbf{y})$. Como mostrado no Exemplo 2 (Subseção 7.2), se $\mu(\mathbf{y}) = \mathrm{M}\mathbf{y}$, então $\widehat{\mathrm{df}} = \operatorname{div} \mu(\mathbf{y}) = tr(\mathrm{M}) = \mathrm{p}$ e neste caso simples o estimador não possui variabilidade e é igual ao df.

No lema de Stein, tem-se uma hipótese matemática bastante técnica que o modelo $\mathrm{M} : \mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}$, em que $\mathrm{M}_{\mathrm{i}}(\mathbf{y}) = (\mu(\mathbf{y}))_{\mathrm{i}}$, tem que ser quasidiferenciável. Tal fato torna o cálculo do $\operatorname{div}\mu(\mathbf{y})$, para os casos em questão (LASSO, *Elastic Net*, Ridge) bastante complexos. Neste trabalho apenas alguns aspectos serão abordados.

7.5 Grau de liberdade na estimação por encolhimento

Esta subseção está baseada no artigo de Kato (2009). No contexto geral do processo de estimação, $\mathrm{K_p}$ é um convexo fechado no espaço paramétrico e o estimador é

$$\hat{\boldsymbol{\beta}}_{\mathrm{K_p}}(\mathbf{y}) = \min_{\boldsymbol{\beta} \in \mathrm{K_p}} \left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y}) \right\|_{\mathrm{p}},$$

com $\hat{\boldsymbol{\beta}}^{\mathrm{o}} = \hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, $\mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{K_p}}(\mathbf{y}) = \mu_{\mathrm{K}}(\mathbf{y})$ e $\mu : \mathbb{R}^{\mathrm{n}} \mapsto \mathrm{K}$ é a projeção de distância mínima no convexo $\mathrm{K} = \mathrm{X}\,\mathrm{K_p}$.

O primeiro resultado, conforme em Kato (2009, p. 4), é bastante técnico e garante que $\boldsymbol{\mu}_{\mathrm{K}}$ e $\boldsymbol{\beta}_{\mathrm{K_p}}$ são funções absolutamente contínuas e, portanto, satisfazem a condição do lema de Stein. Pelo Lema de Stein, um estimador não viesado para o grau de liberdade é o divergente de $\hat{\mu}_{\mathrm{K}}$ (o divergente é relativo à variável $\mathbf{y}$). Segue agora uma relação entre os divergentes das funções μ_{K} e $\beta_{\mathrm{K_p}}$, em que o divergente de $\hat{\beta}_{\mathrm{K_p}}$ é em relação à variável β.

Seja $\mathrm{P}_{\mathrm{K_p}} : \mathbb{R}^{\mathrm{p}} \mapsto \mathrm{K_p}$ a projeção em $\mathrm{K_p}$ de distância mínima em relação à métrica $< \ , \ >_{\mathrm{p}}$ e $\hat{\boldsymbol{\beta}}^{\mathrm{o}} : \mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}^{\mathrm{p}}$, com $\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y}) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ o estimador de quadrados mínimos. Logo, $\hat{\boldsymbol{\beta}}_{\mathrm{K_p}}(\mathbf{y}) = \mathrm{P}_{\mathrm{K_p}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})\right)$ e $\boldsymbol{\mu}_{\mathrm{K}}(\mathbf{y}) = \mathrm{X}\left(\mathrm{P}_{\mathrm{K_p}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})\right)\right)$

e pode-se então utilizar a regra da cadeia.

Se $\mathrm{f} : \mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}^{\mathrm{n}}$ então $\mathrm{df}\,(\mathrm{x})$ é uma transformação linear $\mathrm{df}\,(\mathrm{x}) : \mathbb{R}^{\mathrm{n}} \mapsto \mathbb{R}^{\mathrm{p}}$. A definição de divergente independente de coordenadas é $\mathrm{div\,f}\,(\mathrm{x}) = \mathrm{tr}\,(\mathrm{df}\,(\mathrm{x}))$. Se f é uma transformação linear $\mathrm{f}\,(\mathbf{y}) = \mathrm{M}\mathbf{y}$ então $\mathrm{df}\,(\mathbf{y}) = \mathrm{M}$ e $\mathrm{div\,f}\,(\mathbf{y}) = \mathrm{tr}\,(\mathrm{M})$. No caso de composição, $\boldsymbol{\beta} : \mathbb{R}^{\mathrm{n}} \stackrel{f}{\longmapsto} \mathbb{R}^{\mathrm{p}} \stackrel{g}{\longmapsto} \mathbb{R}^{\mathrm{n}}$, a derivada é dada pela composição de transformações lineares $\mathrm{d}\,(\mathrm{g} \circ \mathrm{f})\,(\mathbf{x}) = \mathrm{d\,g}\,(\mathrm{f}\,(\mathbf{x})) \cdot \mathrm{df}\,(\mathbf{x})$. Tem-se então

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{K}}(\mathbf{y}) &= \mathrm{X}\hat{\boldsymbol{\beta}}_{\mathrm{K_p}}(\mathbf{y}) \\ &= \mathrm{X}\mathrm{P}_{\mathrm{K_p}}\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y}) \\ &= \mathrm{X}\mathrm{P}_{\mathrm{K_p}}\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathbf{y}. \end{aligned}$$

Logo, $\mathrm{d}\,\boldsymbol{\mu}_{\mathrm{K}}(\mathbf{y}) = \mathrm{X} \cdot \mathrm{d\,P}_{\mathrm{K_p}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})\right)(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$. Segue então que

$$\begin{aligned} \mathrm{div}\,\boldsymbol{\mu}_{\mathrm{K}}(\mathbf{y}) &= \mathrm{tr}\left(\mathrm{X\,dP}_{\mathrm{K_p}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})\right)\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\right) \\ &= \mathrm{tr}\left(\mathrm{d\,P}_{\mathrm{K_p}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})\right)\left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'\mathrm{X}\right) \\ &= tr\left(\mathrm{d\,P}_{\mathrm{K_p}}\left(\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})\right)\right). \end{aligned}$$

Portanto o divergente de $\boldsymbol{\mu}_{\mathrm{K}}$ no ponto $\mathbf{y}$ é igual ao divergente da função $\mathrm{P}_{\mathrm{K_p}} : \mathbb{R}^{\mathrm{p}} \mapsto \mathrm{K_p} \subset \mathbb{R}^{\mathrm{p}}$, em relação à variável $\boldsymbol{\beta}$, aplicado no ponto $\hat{\boldsymbol{\beta}}^{\mathrm{o}}(\mathbf{y})$.

Como exemplo, suponha $\mathrm{K_p}$ uma bola de raio r centrada na origem. Para um ponto $\boldsymbol{\beta}$ fora da bola, $\mathrm{P}_{\mathrm{K_p}}(\boldsymbol{\beta}) = \mathrm{r}\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||}$. Logo, ao se derivar em relação a β_{i} tem-se

$$\frac{\partial}{\partial\beta_{\mathrm{i}}}\left(\frac{\mathrm{r}\beta_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}}\right) = \mathrm{r}\frac{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2} - \frac{1}{2}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2\right)^{-\frac{1}{2}}(2\beta_{\mathrm{i}})\,\beta_{\mathrm{i}}}{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}.$$

Como o $\mathrm{div}\left(\mathrm{P}_{\mathrm{K_p}}(\boldsymbol{\beta})\right)$ é o somatório, em i, de tal derivada, então

$$\begin{aligned}
\mathrm{div}\left(\mathrm{P}_{\mathrm{K_p}}(\boldsymbol{\beta})\right) &= \sum_{\mathrm{i}=1}^{n}\left(\frac{\mathrm{r}\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}-\frac{1}{2}\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2\right)^{-\frac{1}{2}}(2\beta_{\mathrm{i}})\,\beta_{\mathrm{i}}}{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}\right)\\
&= \frac{\mathrm{r}\left[\mathrm{n}\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}-\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2\right)^{-\frac{1}{2}}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\beta_{\mathrm{i}}^2\right)\right]}{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}\\
&= \frac{\mathrm{r}\left[\mathrm{n}\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}-\left(\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2\right)^{\frac{1}{2}}\right]}{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}\\
&= \frac{\mathrm{r}\left[(\mathrm{n}-1)\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}\right]}{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}\\
&= \frac{\mathrm{r}(\mathrm{n}-1)}{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}}\beta_{\mathrm{j}}^2}}\\
&= \frac{\mathrm{r}(\mathrm{n}-1)}{\|\boldsymbol{\beta}\|}.
\end{aligned}$$

7.6 Graus de liberdade e estimativas Cp

Como os métodos de estimação LARS, LASSO e Elasticnet são também métodos de seleção de covariáveis é necessário algum critério de seleção de modelos encaixados para a conclusão do processo de estimação. Uma das possibilidades é a de se utilizar estatísticas tipo Cp de Mallows.

Conforme demonstrado anteriormente, se $\hat{\mu} = \mathrm{g}(\mathbf{y})$ então

$$\mathrm{E}\left[\frac{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|}{\sigma^2}\right] = \mathrm{E}\left[\frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\sigma^2}\right] - n + 2\frac{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{cov}(\hat{\mu}_\mathrm{i}, \mathrm{y}_\mathrm{i})}{\sigma^2},$$

em que o grau de liberdade do estimador é dado por $\mathrm{df}_{\boldsymbol{\mu},\sigma^2} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{cov}(\hat{\mu}_\mathrm{i}, \mathrm{y}_\mathrm{i})$ e a estatística Cp fica da forma $\mathrm{Cp}(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\sigma^2} n + 2\mathrm{df}_{\boldsymbol{\mu},\sigma^2}$.

Como $\boldsymbol{\mu}$, σ^2 e $\mathrm{df}_{\boldsymbol{\mu},\sigma^2}$ são parâmetros populacionais, para que Cp possa ser utilizada é necessário estimar estes parâmetros. É sugerido então (Eron et al., 2004) o procedimento que se segue. Obtém-se os estimadores de quadrados mínimos de $\bar{\boldsymbol{\mu}}$ e $\bar{\sigma}^2$ considerando o modelo completo. Para estimar $\mathrm{df}_{\boldsymbol{\mu},\sigma^2}$ é utilizado o método *bootstrap*. Amostras *bootstrap* $\mathbf{y}^*$ são geradas supondo $\mathbf{y}^* \sim \mathrm{N}(\bar{\mu}, \bar{\sigma}^2)$ e estimativas $\hat{\boldsymbol{\mu}}^* = \mathrm{g}(\mathbf{y}^*)$. Repetindo o processo de forma independente por, digamos, B vezes obtém-se estimativas para as covariâncias

$$\widehat{\mathrm{cov}}_\mathrm{i} = \sum_{\mathrm{b}=1}^{\mathrm{B}} \frac{\hat{\boldsymbol{\mu}}_\mathrm{i}^*(\mathrm{b})\left[\mathbf{y}_\mathrm{i}^*(\mathrm{b}) - \mathbf{y}_\mathrm{i}^*(\cdot)\right]}{\mathrm{B} - 1},$$

em que $\mathbf{y}^*(\cdot) = \frac{\sum_{\mathrm{b}=1}^{\mathrm{B}} \mathbf{y}^*(\mathrm{b})}{\mathrm{B}}$, obtendo-se então $\widehat{\mathrm{df}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\widehat{\mathrm{cov}}_\mathrm{i}}{\sigma^2}$.

A suposição de normalidade parece não ser essencial nos resultados, conforme indica a simulação computacional em Efron et al. (2004). Os resultados obtidos sugerem que o grau de liberdade para o estimador LARS no k-ésimo passo é $\mathrm{df}\left(\hat{\boldsymbol{\mu}}_\mathrm{k}\right) \approx \mathrm{k}$, e portanto igual ao grau de liberdade do estimador OLS com k covariáveis. Neste caso, a estimativa para a seleção dos modelos fica da forma

$$\mathrm{C_p} = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_\mathrm{k}\|^2}{\bar{\sigma}^2} - \mathrm{n} + 2\mathrm{k}.$$

O resultado computacional de fato revela uma propriedade verdadeira.

Teorema 3. *Se as covariáveis* $\mathbf{x}_1, \ldots, \mathbf{x}_{\mathrm{m}}$ *são ortogonais, então o estimador* $\hat{\boldsymbol{\mu}}_{\mathrm{k}}(\mathbf{y})$ *dado pelo k-ésimo passo do algoritmo LARS possui* $\mathrm{df}\,(\hat{\boldsymbol{\mu}}_{\mathrm{k}}) = \mathrm{k}$.

Demonstração. Do Lema 3 do artigo, cuja demonstração será omitida, tem-se que $\hat{\boldsymbol{\mu}}_{\mathrm{k}}$ é *quasi diferenciável* e portanto se pode aplicar o Lema de Stein. Assim,

$$\nabla \cdot \hat{\boldsymbol{\mu}}_{\mathrm{k}} = \sum_{\mathrm{i}} \frac{\partial \hat{\mu}_{\mathrm{k,i}}}{\partial \mathbf{y}_{\mathrm{i}}}(\mathbf{y}) = \sum_{\mathrm{i}} \mathrm{I}\left[|\mathbf{y}_{\mathrm{i}}| > |\mathbf{y}_{\mathrm{i}}|_{(\mathrm{k}+1)}\right] = \mathrm{k},$$

uma vez que existem k coordenadas em que $|\mathbf{y}_{\mathrm{i}}| > |\mathbf{y}_{\mathrm{i}}|_{(\mathrm{k}+1)}$. □

7.7 Grau de liberdade do estimador Ridge

Uma das formas de se determinar o grau de liberdade do estimador Ridge é como se segue.

Sendo $\mu(\mathbf{y}) = \mathrm{X}(\mathrm{X}'\mathrm{X} + \mathrm{kI})^{-1}\mathrm{X}'\mathbf{y}$, pode-se escrever

$$\begin{aligned}
\mathrm{df} &= \mathrm{cov}(\mathbf{y}, \mu(\mathbf{y})) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{cov}(\mathbf{y}_{\mathrm{i}}, \mu(\mathbf{y})_{\mathrm{i}}) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{cov}\left(\mathbf{y}_{\mathrm{i}}, \left(\mathrm{X}\big(\mathrm{X}'\mathrm{X} + \mathrm{kI}\big)^{-1}\mathrm{X}'\mathbf{y}\right)_{\mathrm{i}}\right) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{cov}\left(\mathbf{y}_{\mathrm{i}}, \left(\mathrm{X}\big(\mathrm{X}'\mathrm{X} + \mathrm{kI}\big)^{-1}\mathrm{X}'\right)_{\mathrm{i}}\mathbf{y}\right) \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{X}\big(\mathrm{X}'\mathrm{X} + \mathrm{kI}\big)^{-1}\mathrm{X}'\right)_{\mathrm{ii}} \mathrm{cov}\left(\mathbf{y}_{\mathrm{i}}, \mathbf{y}_{\mathrm{i}}\right) \\
&= \mathrm{tr}\left(\mathrm{X}\big(\mathrm{X}'\mathrm{X} + \mathrm{kI}\big)^{-1}\mathrm{X}'\right)\sigma^2 \\
&= \sigma^2 \mathrm{tr}\left[\big(\mathrm{X}'\mathrm{X} + \mathrm{kI}\big)^{-1}\mathrm{X}'\mathrm{X}\right] \\
&= \sigma^2 \mathrm{tr}\Big[\big(\mathrm{X}'\mathrm{X}\big)^{-1}\big(\mathrm{X}'\mathrm{X} + \mathrm{kI}\big)\Big]^{-1} \\
&= \sigma^2 \mathrm{tr}\Big[\mathrm{I} + \mathrm{k}\big(\mathrm{X}'\mathrm{X}\big)^{-1}\Big]^{-1}.
\end{aligned}$$

Como o traço é invariante por conjugação, isto é, $\mathrm{tr}\left(\mathrm{ABA}^{-1}\right) = \mathrm{tr}\left(\mathrm{B}\right)$, pode-se considerar $\mathrm{X}'\mathrm{X}$ na forma diagonal $\mathrm{X}'\mathrm{X} = \mathrm{diag}(\lambda_\mathrm{i})$. Logo,

$$\begin{aligned}
\mathrm{df} &= \sigma^2 \mathrm{tr}\left[\mathrm{I} + \mathrm{k}\,\mathrm{diag}\left(\lambda_\mathrm{i}^{-1}\right)\right]^{-1} \\
&= \sigma^2 \mathrm{tr}\left[\mathrm{diag}\left(1+\frac{\mathrm{k}}{\lambda_\mathrm{i}}\right)\right]^{-1} \\
&= \sigma^2 \mathrm{tr}\left[\mathrm{diag}\left(\frac{\lambda_\mathrm{i}+\mathrm{k}}{\lambda_\mathrm{i}}\right)\right]^{-1} \\
&= \sigma^2 \sum_{\mathrm{i}=1}^{\mathrm{p}} \frac{\lambda_\mathrm{i}}{\lambda_\mathrm{i}+\mathrm{k}}.
\end{aligned}$$

Observa-se que se $\mathrm{k} = 0$, tem-se o estimador de quadrados mínimos e então $\mathrm{df} = \sigma^2 \sum_{\mathrm{i}=1}^{\mathrm{p}} \frac{\lambda_\mathrm{i}}{\lambda_\mathrm{i}+0} = \mathrm{p}\sigma^2$. Entretanto, para $\mathrm{k} \to \infty$, $\mathrm{df} \to 0$. Desta forma, se tem o comportamento conforme descrito na Figura 54.

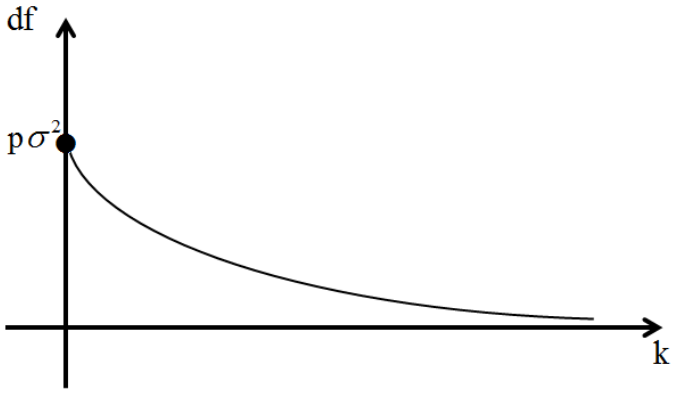


Figura 54 Comportamento do df em relação ao parâmetro Ridge.

7.8 O grau de liberdade do estimador LASSO

Diversos autores (KATO, 2009; ZOU, HASTIE, 2005; TIBSHIRANI, TAYLOR, 2012) demonstram que um estimador não viesado para o grau de liberdade do estimador LASSO é o número de variáveis selecionadas. Este resultado,

como observado pelos autores citados, é surpreendente pois a estimação LASSO se utiliza dos dados para selecionar as variáveis e, aparentemente, o grau de liberdade deverá ser maior que a média do número de variáveis selecionadas. As várias demonstrações deste resultado são bastante teóricas e complicadas. Será então aqui apresentada uma demonstração matematicamente incompleta, porém intuitivamente interessante.

Como visto anteriormente, um estimador não viesado para o grau de liberdade é dado pelo divergente da aplicação que projeta $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}(\mathbf{y})$ em um convexo $\mathrm{K_p}$. Na estimação LASSO este convexo é um poliedro da forma $\sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_{\mathrm{i}}| \leq \mathrm{t}$. Portanto, com alta probabilidade pode-se supor que $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}(\mathrm{t})$ pertence ao interior de uma face l-dimensional do poliedro $\mathrm{K_p}$. Portanto, observando um vetor de dados $\mathbf{y}$ a estimação é simplesmente uma projeção linear em um subespaço afim. Tomando-se como coordenadas de $\mathbb{R}^{\mathrm{p}}$ as coordenadas do plano paralelo à face em que se projeta e as coordenadas restantes em um plano ortogonal, a projeção é dada por uma matriz do tipo da representada na Figura 55, e portanto o divergente é o traço desta matriz, isto é, ℓ, que é o número de componentes de $\hat{\beta}(\mathrm{t})$ diferentes de zero, isto é, o número de covariáveis selecionadas.

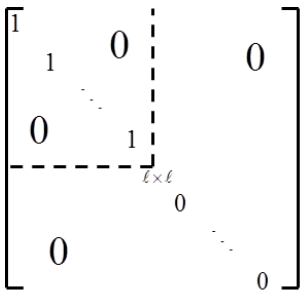


Figura 55 Matriz de projeção.

8 Aspectos computacionais e aplicações

Conforme destacado anteriormente, os métodos LASSO e Elastic Net não possuem forma fechada para se determinar seus respectivos estimadores. Neste sentido, faz-se necessário o emprego de métodos computacionais iterativos para se determinar estimativas de $\boldsymbol{\beta}$. Dentre os vários programas que permitem o cálculo de tais estimativas, optou-se por programas implementados no software R em razão de este ser uma linguagem livre (*freeware*) e por ser amplamente empregado no meio estatístico devido a sua forma simples de programação.

Existem dois principais pacotes no R para aplicação dos métodos LASSO, Ridge e Elastic Net: o elasticnet e o glmnet. Criados respectivamente em 2012 e 2016, ambos são excelentes ferramentas para a computação de estimativas via LASSO/Elastic Net. Entretanto optou-se por utilizar o segundo pacote em detrimento do primeiro, uma vez que nas próprias palavras de um dos autores do pacote (e também criador do método Elasitc Net), Trevor Hastie, o pacote contém procedimentos extremamente eficientes para o ajuste do LASSO ou Elastic Net. Tal eficiência se dá em função da metodologia empregada no algoritmo de minimização da penalização, o qual é baseado em uma metodologia conhecida por *cyclical coordinate descent*, que por sua vez tem o LARS como fundamento. O glmnet é um pacote relativamente simples de ser utilizado. Ainda assim, sua grandiosidade se apresenta principalmente por sua eficiência computacional e por sua robustez, esta última no sentido de que o pacote é ideal para aplicação em *big data*, ou seja, grande quantidade de dados. Segundo Efron et al. (2003), o pacote suportou com alta eficiência conjuntos de dados da ordem de 800 mil covariáveis. Desta forma, os métodos LASSO e Elastic Net, junto com o pacote glmnet, se apresentam como uma ferramenta poderosa para análise de dados genéticos, dada a grande quantidade de dados presentes em amostras genéticas (mesmo com um número pequeno

de observações, geralmente se tem uma informação genética muito grande em tais amostras, isto é, $\mathrm{p} > \mathrm{n}$). Esta situação é exatamente a mesma descrita anteriormente neste texto em que se tem $\mathrm{p} \gg \mathrm{n}$. Segue então uma descrição sobre o pacote a ser utilizado.

O pacote glmnet foi originalmente criado com o intuito de prover uma ferramenta de ajuste de modelo generalizado via máxima verossimilhança penalizada. Mais especificamente fornece, em linguagem R, uma forma de se determinar a estimativa do vetor de parâmetros pelos métodos Ridge, LASSO e Elastic Net. Seus autores, a se saber Friedman, F.; Hastie, T.; Simon, N.; Tibshirani, R., são em sua maioria os próprios autores dos artigos originais dos métodos LASSO e Elastic Net. O comando $\mathrm{glmnet}\,(\mathrm{arguments})$ cria um objeto do tipo glmnet que possui um vasto conjunto de informações referentes ao ajuste realizado. Dois principais (e essenciais) argumentos a serem definidos para a aplicação do pacote são, respectivamente, a matriz das covariáveis (por exemplo X) e um vetor resposta (por exemplo $\mathbf{y}$), sendo que o número de linhas da matriz X deve coincidir com o número de linhas do vetor $\mathbf{y}$. Portanto, uma simples aplicação do pacote seria da forma

```
reg=glmnet(X,y)
```

É importante frisar que com o comando anterior o ajuste, bem como vários parâmetros inerentes a este, é armazenado no objeto reg. Vários outros parâmetros possíveis de se determinar como argumento do pacote são bastante específicos. Desta forma seguem destacados apenas os mais relevantes a nossa pesquisa.

$\mathrm{alpha} = \mathrm{c}$, com $\mathrm{c} \in [0, 1]$, determina se o ajuste a ser feito é do tipo Ridge, LASSO ou Elastic Net. Se $\mathrm{c} = 0$ o pacote executa um ajuste Ridge. Se $\mathrm{c} = 1$, o ajuste realizado é pelo método LASSO. Já para $0 < \mathrm{c} < 1$, é ajustado um modelo do tipo Elastic Net, sendo que quanto mais próximo de 0 o parâmetro c estiver,

mais próximo de uma restrição Ridge será o formato da penalização, enquanto que um valor de c mais próximo de 1 resulta em uma penalização com formato mais próximo a restrição de norma $\| \ \|_1$. Portanto, um comando do tipo

```
reg=glmnet(X,y,alpha=1)
```

irá realizar um ajuste LASSO com o vetor de respostas **y** e com a matriz X, cujas colunas são as covariáveis analisadas. Já um comando do tipo

```
reg=glmnet(X,y,alpha=0.5)
```

irá ajustar um modelo no formato Elastic Net. A função plot(arguments) permite que se visualize os traços gráficos das componentes do vetor estimado para cada valor da norma L1, lembrando que tal norma está intrinsecamente ligada ao parâmetro s da regressão, que por sua vez representa a distância da origem até o extremo do convexo de restrição no espaço paramétrico. Como exemplo, a Figura 56 representa um gráfico feito com um conjunto de dados de 20 covariáveis. Ao

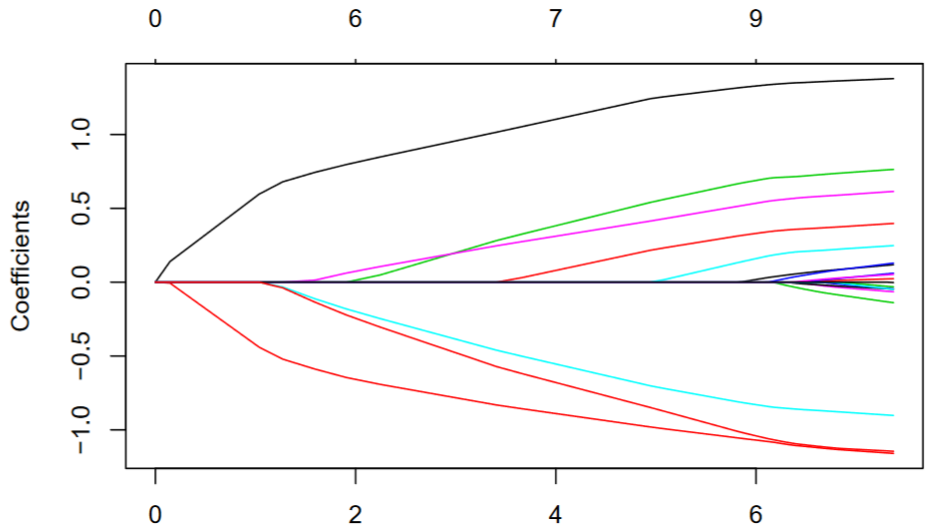


Figura 56 Estimação de parâmetros *Elastic Net*.

se incluir na função plot o argumento label = TRUE, do lado esquerdo do gráfico plotado irão aparecer legendas representando cada uma das covariáveis. Na Figura 57 é possível ver no lado direito da imagem os números que indicam a qual covariável cada uma das linhas traçadas representa. Uma vez realizado o ajuste, é

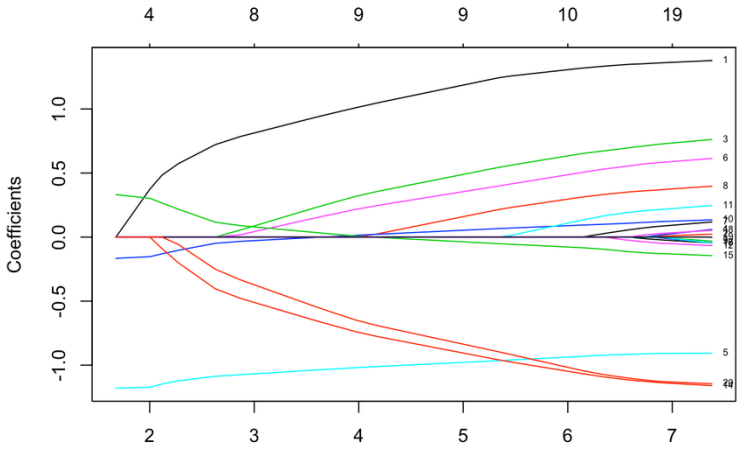


Figura 57 Exemplo de um plot com legenda.

possível se determinar uma estimativa específica dado um valor de s. Assim, ao se digitar

```
coef(reg,s=0.1)
```

temos como resultado uma estimativa pontual do vetor $\boldsymbol{\beta}$, a qual foi extraída do ajuste reg. Uma pergunta natural é então como se determinar qual dentre as várias estimativas dadas por um ajuste é a mais indicada para a escolha do modelo ajustado? O pacote fornece tal estimativa baseado no conceito de *Generalized Cross Validation* (GCV), indicando qual das estimativas resulta em um modelo com Erro Quadrático Preditivo menor, no sentido generalizado. Para tanto, é necessário a criação de um novo objeto, do tipo cv.glmnet. Tal objeto armazena informações sobre o processo de Cross-Validation executado pelo pacote. Assim, para se de-

terminar o vetor estimativa com menor erro preditivo, segundo o GCV, deve-se executar os comandos da forma

```
cvreg = cv.glmnet(X, y,alpha=0.5)
coef(cvreg, s = "lambda.min")
```

Este último comando fornece os coeficientes do vetor estimativa $\hat{\boldsymbol{\beta}}$ com menor GCV, que é então utilizado como vetor de coeficientes para o modelo selecionado.

Como aplicação em dados reais, foram escolhidos dados genéticos referentes a mensuração de pH carne de suínos *Sus Scrofa* 45 min e 24 h após o abate. Foram então analisadas as covariáveis genéticas em termos de regressão Elastic Net, de forma a se agrupar covariáveis correlacionadas e se determinar, através do método, aquelas mais relevantes à variável resposta pH. Um estudo semelhante, porém utilizando *Partial Least Squares* (PLS), sobre os mesmos dados pode ser encontrado em Silveira et al. (2017). Mais detalhes referentes aos dados serão apresentados nas subseções seguintes.

8.1 Simulações no R

Foram realizadas simulações para se testar a aplicabilidade do pacote. Como exemplificação, uma primeira simulação foi feita aos moldes de um exemplo idealizado no artigo original (ZOU; HASTIE 2005, p. 313). Neste exemplo, o objetivo principal é mostrar a diferença básica entre os métodos LASSO e Elastic Net. Foram então criadas duas variáveis aleatórias independentes Z_1 e Z_2, ambas com uma distribuição $\mathrm{U}(0, 20)$. A resposta $\mathbf{y}$ foi gerada por uma $\mathrm{N}(\mathrm{Z}_1 + 0.1\mathrm{Z}_2, 1)$. É importante destacar que, na resposta $\mathbf{y}$, os pesos das variáveis aleatórias são bem distintos, já que a variável Z_2 tem peso dez vezes menor do que a variável Z_1, ou seja, Z_2 contribui bem menos para a resposta. Suponha que foram observadas seis covariáveis "latentes" $\mathbf{x}_1, \ldots, \mathbf{x}_6$ identicamente distribuídas da forma que, para

erros ε_i independentes e identicamente distribuídos por uma $\mathrm{N}(0, 1/16)$ se tem a relação

$$\mathbf{x}_1 = \mathbf{Z}_1 + \varepsilon_1, \qquad \mathbf{x}_2 = -\mathbf{Z}_1 + \varepsilon_2, \qquad \mathbf{x}_3 = \mathbf{Z}_1 + \varepsilon_3$$

$$\mathbf{x}_4 = \mathbf{Z}_2 + \varepsilon_4, \qquad \mathbf{x}_5 = -\mathbf{Z}_2 + \varepsilon_5, \qquad \mathbf{x}_6 = \mathbf{Z}_2 + \varepsilon_6.$$

Claramente, as covariáveis $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ formam um grupo cujo fator subjacente é a variável $\mathbf{Z}_1$, enquanto $\mathbf{x}_4$, $\mathbf{x}_5$ e $\mathbf{x}_6$ possuem $\mathbf{Z}_2$ como fator subjacente. A Figura 58 (ZOU; HASTIE 2005, p. 311) apresenta os traços do LASSO (a) e do Elastic Net (b). É possível ver que, apesar de o LASSO ter o efeito de seleção de covariáveis (pois as linhas do traço surgem de forma sequencial, ao contrário do Ridge em que estas entram simultaneamente), cada linha (relativa a uma covariável) entra separada das demais. Já no traço do Elastic Net, é possível perceber claramente o efeito de agrupamento. Além de os traços das covariáveis $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ surgirem no mesmo momento, o método ainda as apresenta antes das três últimas covariáveis, ou seja, para um valor de s menor. Isso se deve ao fato de que as três primeiras covariáveis estão relacionadas à variável $\mathbf{Z}_1$, que possui uma influência na resposta 10 vezes maior do que a variável $\mathbf{Z}_2$. Além do mais, pode-se perceber no Elastic Net trace que a linha referente à variável $\mathbf{x}_2$ aparece de forma decrescente, bem como a linha referente à variável $\mathbf{x}_5$. Isto ocorre devido a natureza destas duas variáveis, a qual é influenciada de forma negativa pelas covariáveis $\mathbf{Z}_1$ e $\mathbf{Z}_2$, respectivamente. Nota-se que o LASSO, para esse exemplo específico, não foi capaz de perceber esta natureza decrescente referente a variável $\mathbf{x}_5$.

Em resumo o exemplo anterior, apesar de hipotético, ilustra de forma clara as características do método Elastic Net em relação ao LASSO. O efeito de agrupamento, dependendo do tipo de aplicação, pode se apresentar como uma ferramenta crucial e inovadora na análise de dados em que se tem uma gama muito grande de

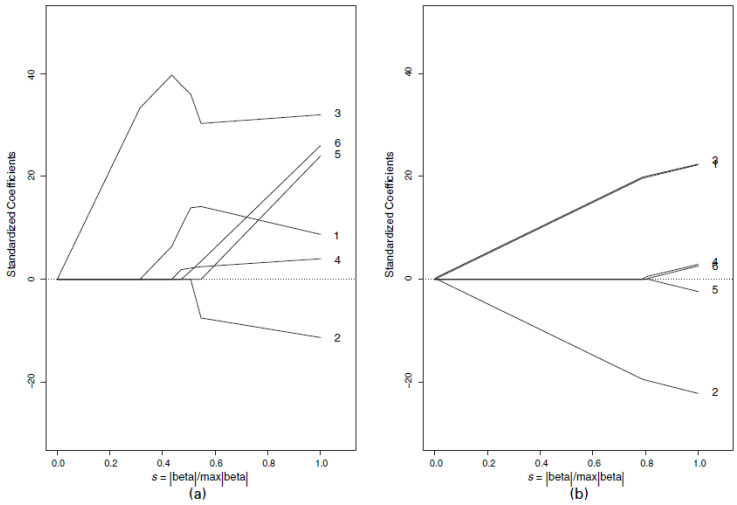


Figura 58 Comparação entre LASSO Trace e Elastic Net Trace.

covariáveis.

O código em R da simulação anterior se encontra no Apêndice na parte referente ao **Código 1**.

Uma segunda simulação (**Código 2** no Apêndice) foi feita no sentido de verificar os efeitos do Elastic Net não em variáveis latentes, mas sim diretamente nas covariáveis resposta. Foram então criadas 5 covariáveis, de forma que a variável $\mathbf{x}_1$ tenha distribuição $\mathrm{N}(0,1)$. A covariável $\mathbf{x}_2$ também é normal, porém com média $\mathbf{x}_1$ e desvio padrão $\mathrm{sd} = 0.1$, ou seja, é altamente correlacionado com a variável $\mathbf{x}_1$. $\mathbf{x}_3$, independente de $\mathbf{x}_1$ e $\mathbf{x}_2$, também segue uma normal $\mathrm{N}(0,1)$ e $\mathbf{x}_4$, também normal, possui média $\mathbf{x}_3$ e desvio padrão de $\mathrm{sd} = 0.008$. Apenas $\mathbf{x}_5$ foi criada de forma a depender de duas das covariáveis anteriores, seguindo então uma normal $\mathrm{N}(\mathbf{x}_4+\mathbf{x}_3, 0.005)$. O vetor resposta $\mathbf{y}$ segue uma distribuição normal com desvio padrão 1, porém duas situações foram analisadas: uma em que $\mathbf{y}$ possui média $0.01\mathbf{x}_1 + \mathbf{x}_2 + 2\mathbf{x}_3 - \mathbf{x}_4 + 2\mathbf{x}_5$ e outra em que se altera o peso da variável $\mathbf{x}_4$ na média de $\mathbf{y}$ de -1 para -10. As Figuras 59 e 60 apresentam, respectivamente,

o traço do Elastic Net para as duas situações. Comparando ambas as situações por

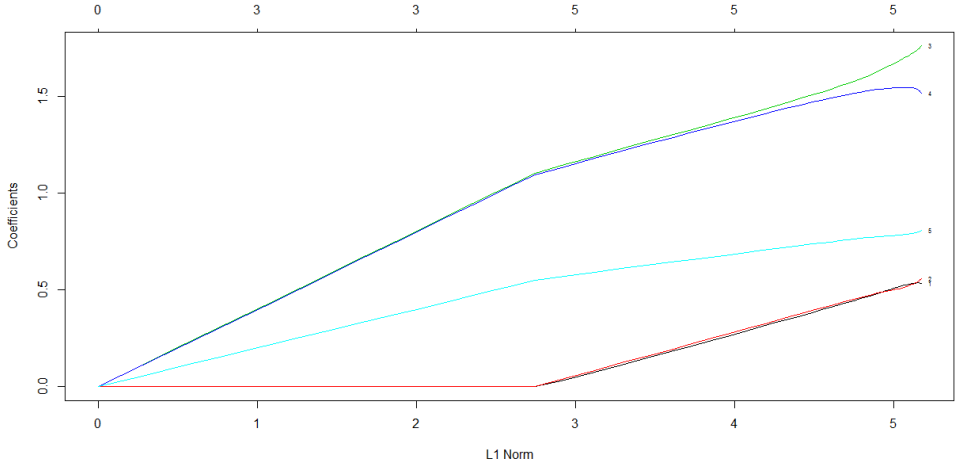


Figura 59 Elastic Net trace para a situação 1.

seus respectivos gráficos, ao se aplicar um peso relativamente grande na covariável $\mathbf{x}_4$, as outras covariáveis correlacionadas a esta ($\mathbf{x}_3$ e $\mathbf{x}_5$), mesmo tendo coeficientes positivos, apresentam comportamento similar ao da covariável de coeficiente negativo ($\mathbf{x}_4$). Tal fato demonstra como o efeito de agrupamento no Elastic Net é considerável.

No **Código 3** (Apêndice), foi reproduzido a plotagem de um exemplo clássico, que primeiro surge em Tibshirani (1996), referente a dados de 8 covariáveis relacionadas ao câncer de próstata (*prostate cancer*) e uma variável resposta, **lpsa**, acrônimo de *log prostate specific antigen*. Os dados, presentes no pacote lasso2, estão armazenados em um *dataframe* de nome Prostate. O gráfico da variável resposta em função das 8 covariáveis é apresentado na Figura 61. Conforme se pode observar, as covariáveis com maior peso na resposta são lcavol, svi e lweight. Posteriormente, Zou; Hastie (2005) utilizaram este mesmo conjunto de dados no artigo do Elastic Net para se fazer um comparativo entre os dois métodos.

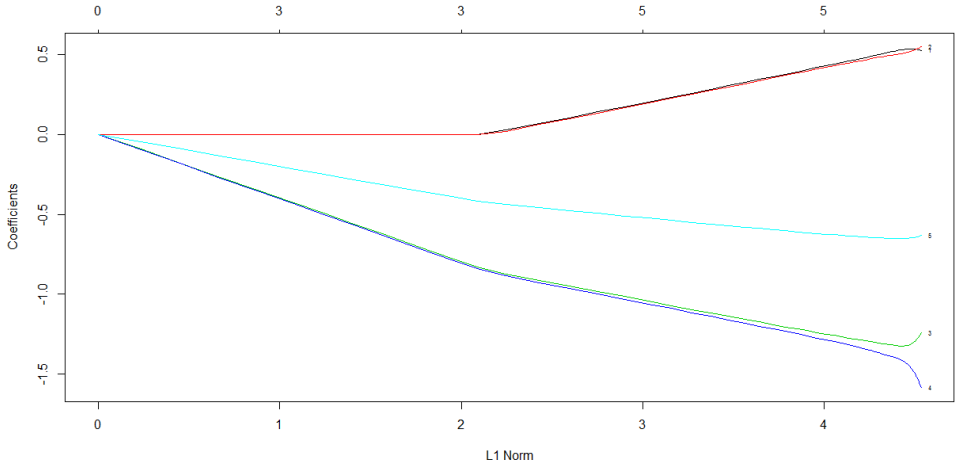


Figura 60 Elastic Net trace para a situação 2.

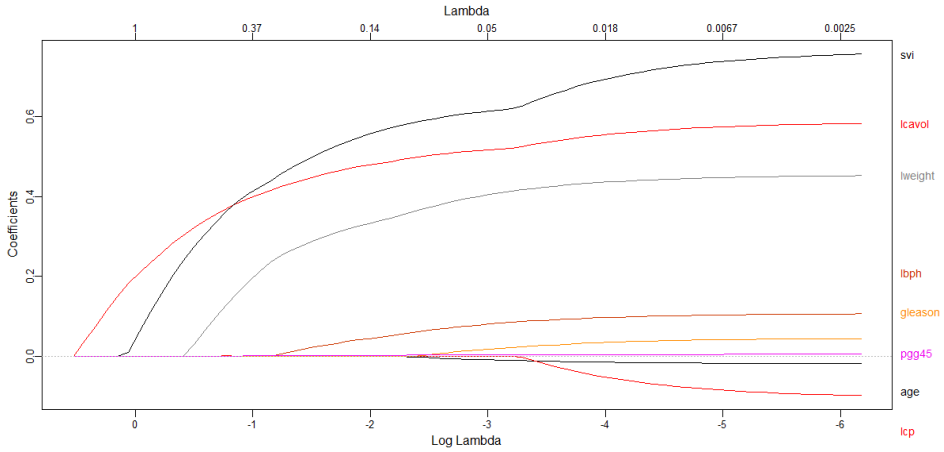


Figura 61 Elastic Net trace para os dados de câncer de próstata.

Uma última simulação (**Código 4** no Apêndice) foi feita com o intuito de se analisar o efeito de covariáveis categóricas no processo dos métodos LASSO e Elastic Net. Tal simulação se faz de interesse uma vez que, em geral, ao se coletar dados genéticos referentes a marcadores moleculares (SNPs), os resultados são

categóricos (heterozigoto, homozigoto dominante e homozigoto recessivo). Um grupo de 8 covariáveis categóricas (com níveis 0, 1 e 2) foi criado por meio de cópulas, de forma que as três primeiras covariáveis formem um grupo correlacionado, as três próximas covariáveis (4, 5 e 6) formem um segundo grupo correlacionado (independente do primeiro) e as duas últimas covariáveis formam um terceiro grupo correlacionado entre si. Em seguida, uma variável resposta foi produzida da forma

$$\mathbf{y} = 130\mathrm{x}_1 + 120.1\mathrm{x}_2 + 150\mathrm{x}_3 - 15\mathrm{x}_4 + 0.2\mathrm{x}_5 - 0.15\mathrm{x}_6 + 60\mathrm{x}_7 - 59\mathrm{x}_8 + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, 0.1)$. O gráfico do Elastic Net trace é conforme Figura 62.

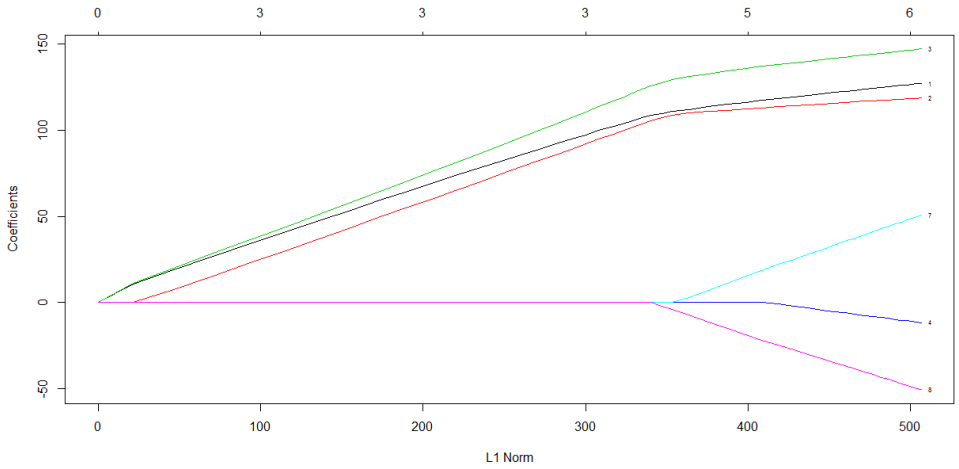


Figura 62 Elastic Net trace para covariáveis categóricos.

Alguns pontos sobre esta última simulação merecem destaque. Primeiramente, e mais importante, é possível dizer que o método Elastic Net (em particular o pacote glmnet) opera normalmente com covariáveis categóricas. Portanto, a aplicação do método em dados categóricos é perfeitamente possível. Novamente, se pode perceber que pela Figura 62 as covariáveis mais relevantes para o vetor

resposta $\mathbf{y}$ surgem primeiro no Elastic Net trace, ou seja, as covariáveis 1, 2 e 3. Além do mais, o efeito de agrupamento se manteve inalterado em covariáveis categóricas. Outro ponto que merece destaque é o fato de, no segundo grupo de covariáveis correlacionadas (4, 5 e 6), apenas a covariável 4 aparece no gráfico. As outras duas covariáveis (5 e 6) não aparecem uma vez que o valor de seus coeficientes (ou pesos) são praticamente desprezíveis em relação aos valores dos demais coeficientes. Fato que corrobora esta última afirmação é o valor da estimativa que resulta no menor erro quadrático, por meios de GCV, já que as estimativas dos coeficientes 5 e 6 são zeradas por este critério. A Figura 63 apresenta os valores estimados por GCV para a simulação em R.

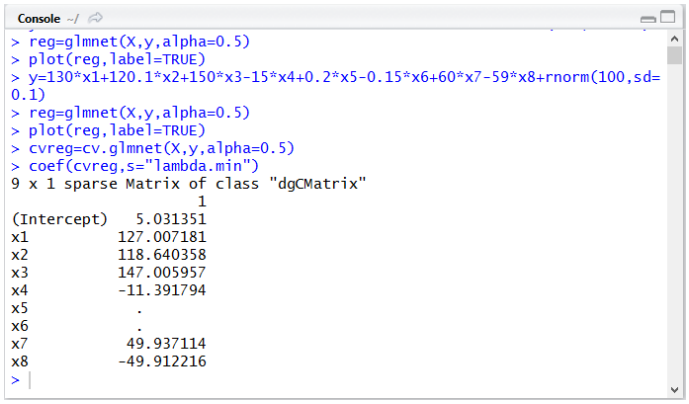


Figura 63 Estimativa do vetor de parâmetros dos dados categóricos.

8.2 Aplicação em dados genéticos

Como aplicação em dados reais, optou-se por adotar dados genéticos referentes a mensuração de pH da carne e peso de carcaça de suínos ***Sus scrofa***. Os dados são constituídos por 345 observações e 300 covariáveis, as quais estão alocadas nas colunas do arquivo dos dados. Entretanto, os 237 marcadores genéticos (SNPs) estão concentrados entre as colunas 64 e 300 ($\mathrm{ALGA0000087} \sim \mathrm{ASGA0080951}$)

do arquivo de dados. Conforme descrito em Silveira et. al. (2017), os 237 marcadores estão distribuídos em seis cromossomos do ***Sus scrofa*** da seguinte maneira: SSC1(56), SSC4(54), SSC7(59), SSC8(30), SSC17(25) e SSCX(13). Apesar de todas as propriedades dos métodos LASSO e Elastic Net, a análise gráfica simultânea das 237 covariáveis se torna um tanto quanto impraticável, conforme apresentado na Figura 64, em que foi escolhida como variável resposta o pH mensurado a 45 minutos após o abate.

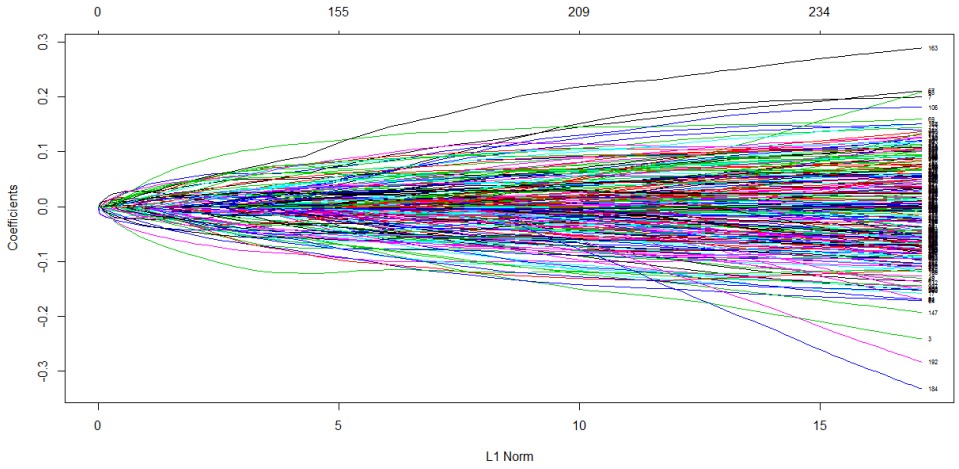


Figura 64 Elastic Net trace das 237 covariáveis.

Desta forma, o código em R foi implementado de modo que os índices j e k selecionem, respectivamente, a primeira e a última coluna da matriz X utilizada na análise, respeitando a relação $1 \leq j < k \leq 237$. As linhas de comando do código constam no Apêndice (**Código 5**). As linhas de comando

```
X<-X[-c(which(is.na(y))),]
y=y[-c(which(is.na(y)))]
```

foram incluídas em função dos valores das respostas, já que algumas observações

possuem o valor NA, provavelmente para representarem parcelas perdidas. Assim, os comandos anteriores excluem tanto do vetor resposta $\mathbf{y}$ quanto da matriz X as linhas cujas entradas relativas ao vetor $\mathbf{y}$ contenham o valor NA. Como referência, a partir de agora os marcadores serão denominados de acordo com sua posição a partir da coluna 64, isto é, o marcador 3 é aquele referente a coluna 66 dos dados brutos.

9 Resultados e discussão

Buscou-se determinar dentre todos os 237 marcadores os 12 (aproximadamente 5%) que mais influenciam na resposta peso de carcaça PCARC. Novamente, a Figura 65 representa a dificuldade de se realizar tal análise graficamente. Para se contornar tal problema foi utilizada a ferramenta que informa o valor de

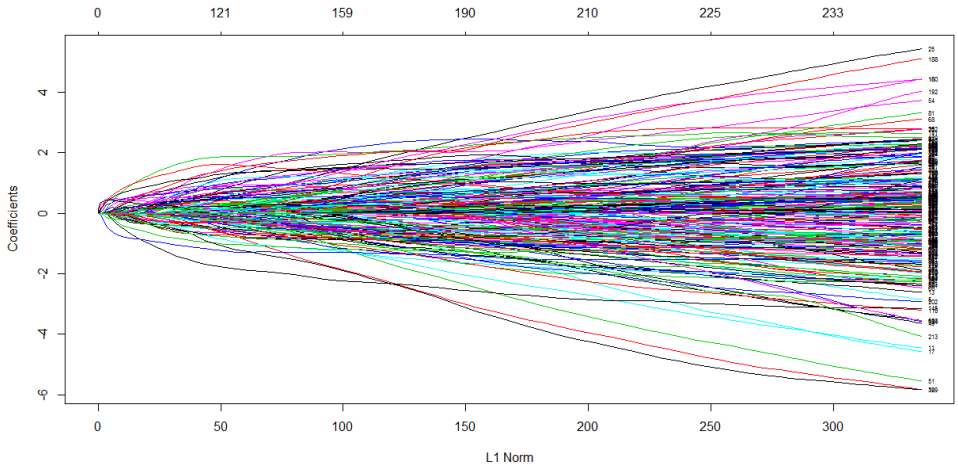


Figura 65 Elastic Net trace para as 237 covariáveis com a resposta PCARC.

uma estimativa específica para um dado valor de s e então, atribuindo valores altos para s (o que implica em valores baixos da norma L1) e fazendo-o decrescer gradativamente, as covariáveis mais importantes surgem também de forma gradativa.

Por esta metodologia, as 12 covariáveis selecionadas foram as de número 36, 82, 129, 144, 148, 162, 190, 206, 208, 211, 219 e 229. Plotando estas covariáveis em um Elastic Net trace tem-se a Figura 66. O código para a execução desta última análise sofreu algumas alterações (em relação ao **Código 5**), as mais relevantes sendo nas linhas que montam a matriz X cujas colunas são formadas pelas 12 covariaveis selecionadas, e também foram excluídos os comandos que retiram

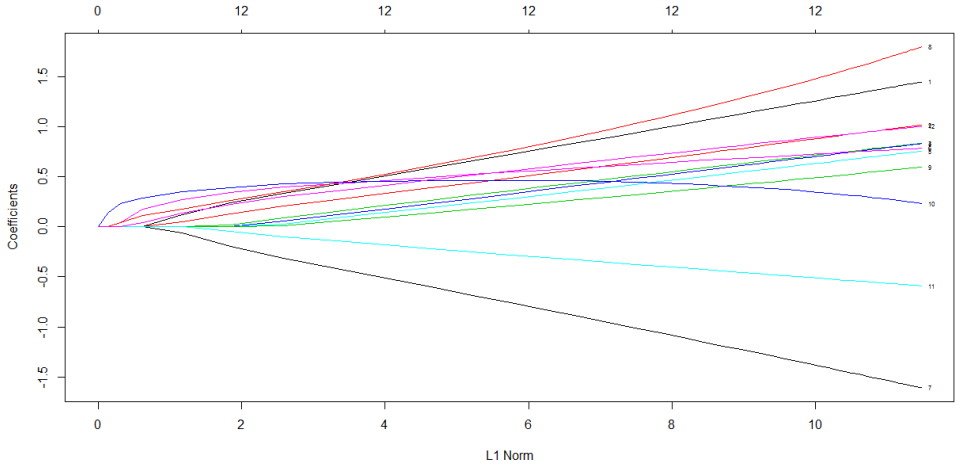


Figura 66 As 12 covariáveis mais relevantes para a resposta PCARC.

linhas da matriz X cujos respectivos elementos em **y** sejam NA, uma vez que utilizando PCARC como resposta, não se tem nenhuma parcela perdida. O código sumarizado está no Apêndice, **Código 6**. A Figura 67 representa de forma esquemática a localização, através das setas vermelhas, de cada uma das 12 covariáveis selecionadas na sequência total dos 220 marcadores, bem como suas posição relativa a cada um dos 6 cromossomos. Uma análise da Figura 67 mostra que a grande

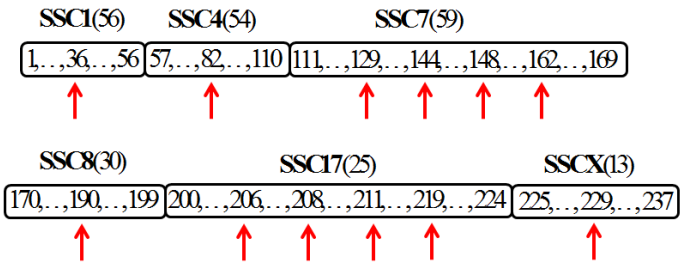


Figura 67 Posição das 12 covariáveis mais relevantes em relação aos grupos de ligação.

maioria dos marcadores (dentre os 12 mais influentes para a resposta PCARC)

estão presentes nos cromossomos SSC7 e SSC17.

Foi realizada também uma análise similar utilizando como resposta o fenótipo pH45. Desta forma, 13 covariáveis (5%) foram selecionadas como sendo mais relevantes dentre as 237, a se saber, 11, 13, 49, 54, 64, 86,126, 133, 154, 157, 186, 199 e 234. As linhas de comando do R constam no Apêndice (**Código 7**). As curvas Elastic Net para tal situação estão representadas na Figura 68.

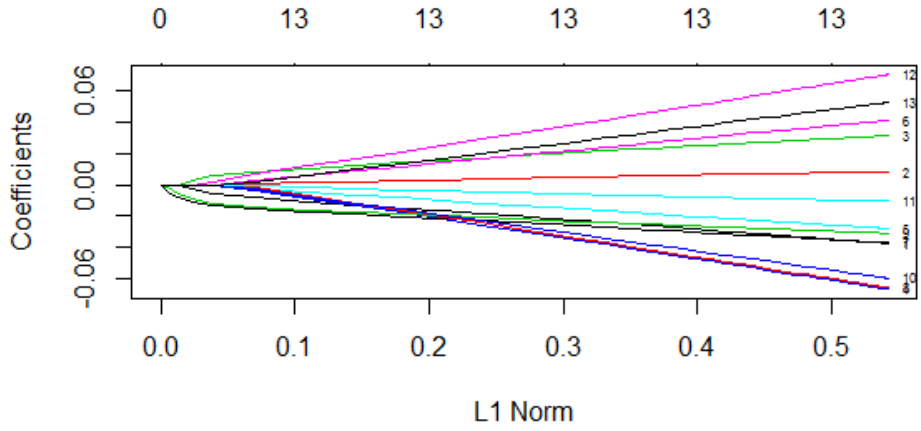


Figura 68 As 13 covariáveis mais relevantes para a resposta PH45.

No sentido de se observar melhor as covariáveis em um cromossomo específico, optou-se por buscar as 9 covariáveis mais relevantes no grupo de ligação SSC4, composto por 54 marcadores genéticos (colunas 57 a 110). O traço Elastic Net para esta situação é conforme Figura 69. Os comandos em R constam no Apêndice (**Código 8**). O cromossomo SSC4 foi escolhido em detrimento dos demais para que possa ser feita uma análise comparativa com os resultados de Silveira et al. (2017). Conforme se pode observar, os 9 marcadores selecionados no grupo de ligação SSC4 foram 64, 67, 69, 72, 76, 81, 86, 87 e 101.

Estes resultados são discrepantes com aqueles obtidos por Silveira, op.

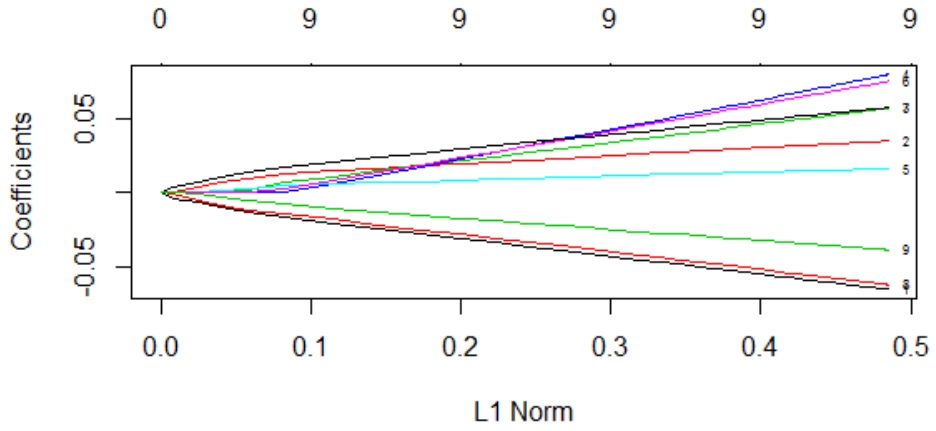


Figura 69 As 9 covariáveis mais relevantes para a resposta PH45 no cromossomo SSC4.

cit., uma vez que o método PLS identificou como mais relevantes os marcadores 62, 94, 98, 91 e 93. Mesmo a análise especifica no cromossomo SSC4 observa-se que os marcadores selecionados pelo método PLS diferem daqueles selecionados pelo método Elastic Net.

Para uma análise mais específica, o método Elastic Net foi aplicado aos marcadores entre 90 e 99 (localizados no cromossomo 4). A Figura 70 apresenta o gráfico Elastic Net para estes 10 marcadores, no o caso em que a variável resposta é o pH 45 minutos após o abate. No eixo das abscissas, foi colocado entre parêntesis o número dos marcadores, de forma a facilitar a visualização e interpretação dos resultados. Conforme se pode observar, os marcadores 98 e 94 possuem maior influência na resposta quando comparado neste grupo de 10 covariáveis, inclusive apresentando um efeito de agrupamento acentuado. Este resultado segue em conformidade com o artigo citado. Em seguida, as covariáveis 96, 93 e 92 formam um segundo grupo correlacionado e apresentam significativa relevância

para o resultado, sendo que o marcador 93 foi identificado como de alta relevância pelo referido artigo. Fica claro também pelo gráfico que existe um agrupamento entre os marcadores 99 e 97, porém o peso destes na resposta é bem menor quando comparado com os demais, uma vez que estes aparecem mais a direita no eixo das abscissas. Um resultado evidenciado pelo método Elastic Net e não captado pelo

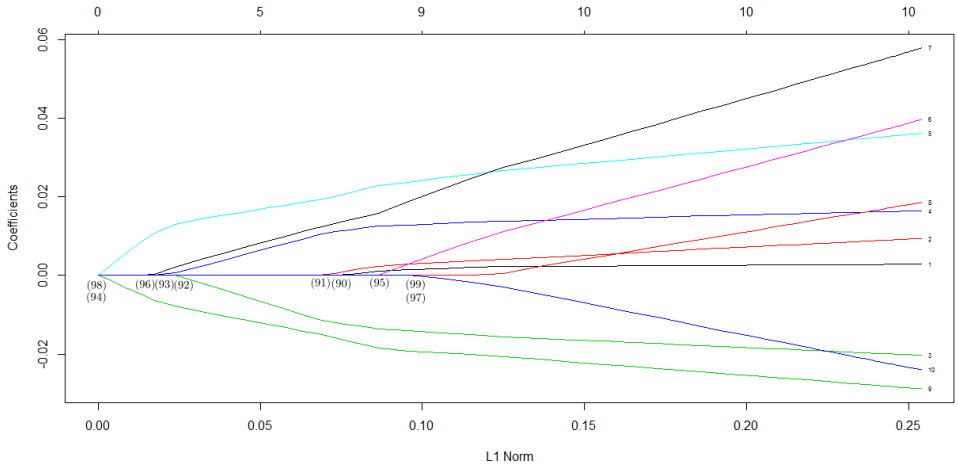


Figura 70 Elastic Net trace para os marcadores de 90 a 99.

método PLS, para estas 10 covariáveis analisadas, é que os marcadores 96 e 92 apresentam efeito similar ao marcador 93, em função do efeito de agrupamento destas três covariáveis. Entretanto, o valor da estimativa da covariável 92 na resposta apresenta valor negativo, enquanto que a estimativa das covariáveis 96 e 93 possuem valores positivos.

O método do Elastic Net, de forma análoga aos exemplos anteriores, pode ser aplicado em qualquer uma das várias respostas contidas no arquivo de dados (PCD, RCARC, PROLOM, PCOPA, etc.). Entretanto, o presente trabalho tomou apenas as respostas pH45 e PCARC para análise, uma vez que o objetivo principal de tais análises é apresentar a aplicabilidade do método Elastic Net, bem

como suas vantagens e características. Com os códigos que constam no Apêndice, uma vasta gama de análises pode ser feita com o conjunto de dados utilizado, bastando apenas pequenas alterações nas linhas de comando apresentadas. O importante a se destacar é que o método Elastic Net se apresenta como uma ferramenta bastante aplicável na análise de dados genéticos, principalmente no que tange a seleção de covariáveis por grupos correlacionados.

10 CONCLUSÃO

- Uma abordagem geométrica à teoria dos estimadores LASSO, LARS e Elastic Net se revelou eficiente no sentido de ser intuitiva e explicitar aspectos essenciais da teoria.

- O texto representa uma contribuição para a leitura dos artigos clássicos da teoria LASSO, LARS e Elastic Net.

- As rotinas desenvolvidas utilizando o pacote glmnet apresentaram bons resultados e podem ser utilizadas para futuras análises em dados genéticos.

- Os resultados obtidos na análise dos dados genéticos de ***Sus scrofa*** apresentam uma relevante similaridade em relação aos resultados de Silveira et al. (2017).

- Novos resultados são apresentados na detecção de SNPs relevantes na variável resposta Peso de Carcaça.

REFERÊNCIAS

BOLDRINI, J. L. et al. **Álgebra Linear**. 2. ed. São Paulo: Harbra, 1986. 411 p.

BREIMAN, L. Better subset regression using the nonnegative garrote. **Technometrics**, 1995, vol. 37, n. 4. p. 373 to 384

COSTA, L. A. Novo estimador de cumeeira de Rao com aplicação em seleção genômica. **Tese (Mestrado em Estatística e Experimentação Agropecuária)** - Universidade Federal de Lavras, Lavras-MG, 2015., p. 126.

EFRON, B; HASTIE, T; JOHNSTONE, I; TIBSHIRANI, R. Least angle regression. **The Annals of Statistics**, 2004, vol. 32, n. 2. p. 407 to 499

EFRON, B. The estimation of prediction error: covariance penalties and cross-validation. **Journal of the American Statistical Association**, 2004, vol. 99, n. 467. p. 619 to 632

GAJO, C. A. Propriedades e aspectos geométricos de estimadores tipo James-Stein e do estimador de Hartigan. **Tese (Doutorado em Estatística e Experimentação Agropecuária)** - Universidade Federal de Lavras, Lavras-MG, 2016, p. 194.

GENTLE, J. E. **Matrix algebra: theory, computations and applications in statistics**. New York: springer, 2007. 528 p.

GOLUB G. H., HEATH, M., WAHBA, G. Generalized cross-validation as a method for choosing a good ridge parameter. **Thecnometrics**, 1979, vol. 21, n. 21. p. 215 to 223

GRUBER, M. H. J. **Improving efficiency by shrinkage: the James-Stein and ridge regression estimator**. 2nd ed. New York: M. Dekker, 1998. 632 p.

HASTIE, T; TIBSHIRANI, R; FRIEDMAN, J. **The elements of statistical learning. Data mining, Inference and Prediction**. New York: Springer, 2nd ed.

2008. 739 p.

HOERL, A.E; KENNARD, R. W. Ridge regression: Applications to nonorthogonal problems. **Thecnometrics**, 1970, vol. 12, n. 1. p. 69 to 82

HOERL, A.E; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Thecnometrics**, 1970, vol. 12, n. 1. p. 51 to 67

JAMES, G; WITTEN, D; HASTIE, T; TIBSHIRANI, R. **An introduction to statistical learning with application in R**. New York: Springer, 2013. 426 p.

KATO, K. On the degrees of freedom in shrinkage estimation. **Journal of Multivariate Analysis**, v. 100, p. 1338-1352, 2009.

PEREIRA, L. S. Abordagem geométrica à teoria dos modelos de Gausss-Markov. **Dissertação (Mestrado em Estatística e Experimentação Agropecuária)** - Universidade Federal de Lavras, Lavras-MG, 2013, p. 130.

PEREIRA, L. S.; CHAVES, L. M, et al.; Geometry of Basic Properties on Linear Regression and The Mallow's Cp Statistics. **Revista Brasileira de Biometria**, v. 33, n. 3, p. 357-377, 2015.

RENCHER, A. C; SCHAALJE, G.B. **Linear Models in statistics**. New Jersey: John Wiley & Sons, 2008. 672 p.

SAVILLE, D. J; WOOD, G. L. A Method for Teaching Statistics Using N-Dimensional Geometry. **The American Statistician**, v. 40, p. 205-214, 1986.

SAVILLE, D. J; WOOD, G. L. **Statistical Methods: The Geometric Aproach**. New York: Springer-Verlag, 1991. 560 p.

SILVEIRA, F. G; COSTA, L. A.; PEREIRA, L. S, et al.; Uma demonstração geométrica para uma identidade de Fisher para o modelo de dois fatores. **Revista Brasileira de Biometria**, v. 30, n. 2, p. 199-222, 2012.

SILVEIRA, F. G; DUARTE, D. A. S.; CHAVES, L. M; SILVA, F. F; FILHO, I. C; DUARTE, M. S; LOPES, P. S; GUIMARÃES, S. E. F; The optimal number of partial least squares components in genomic selection for pork pH **Ciência Rural, Santa Maria**, v. 47, No.1, p. 1-5, 2017.

TIBSHIRANI, R. J; TAYLOR, J. Degrees of freedom in lasso problems. **The Annals of Statistics**, v. 40, No.2, p. 1198-1232, 2012.

TIBSHIRANI, R. Regression Shrinkage and Selection via the LASSO. **Journal of the Royal Statistical Society**, v. 58, No.1, p. 267-288, 1996.

ZOU, H; HASTIE, Y. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society**, v. 67, No.2, p. 301-320, 2005.

APÊNDICE

Código 1

```
library(glmnet)
z1=runif(100, min = 0, max = 20)
z2=runif(100, min = 0, max = 20)
x1=z1+rnorm(100,mean=0,sd=1/16)
x2=-z1+rnorm(100,mean=0,sd=1/16)
x3=z1+rnorm(100,mean=0,sd=1/16)
x4=z2+rnorm(100,mean=0,sd=1/16)
x5=-z2+rnorm(100,mean=0,sd=1/16)
x6=z2+rnorm(100,mean=0,sd=(1/16))
u<-cbind(x1,x2,x3,x4,x5,x6)
y=rnorm(100,mean=z1+0.1*z2,sd=1)
reg=glmnet(u,y,alpha=0.5)
plot(reg,label=TRUE)#ELASTICNET
reg=glmnet(u,y,alpha=1)
plot(reg,label=TRUE)#LASSO
```

Código 2

```
library(glmnet)
rm(list=ls())
x1<-rnorm(100)
x2<-rnorm(100,mean=x1,sd=.01)
x3<-rnorm(100)
x4<-rnorm(100,mean=x3,sd=.008)
x5<-rnorm(100,mean=(x4+x3),sd=.005)
```

```
u<-cbind(x1,x2,x3,x4,x5)
##########################################
y<-rnorm(100,mean=0.01*x1+x2+2*x3-x4+2*x5)
reg=glmnet(u,y,alpha=0.5)
plot(reg,label=TRUE)
##########################################
y<-rnorm(100,mean=0.01*x1+x2+2*x3-10*x4+2*x5)
reg=glmnet(u,y,alpha=0.5)
plot(reg,label=TRUE)
```

Código 3

```
install.packages("lasso2")
library(lasso2)#dados do Prostate Cancer
library(glmnet)
rm(list=ls())
prost=data(Prostate)
ls(Prostate) #conteudo  do Prostate Data
X = as.matrix(Prostate[setdiff(colnames(Prostate),
"lpsa")])
#o comando anterior retira "lpsa" dos dados e monta
#a matriz X com os demais
y<-Prostate$lpsa
reg=glmnet(X,y,alpha=0.5)
plot(reg,label=TRUE)
```

Código 4

```
library(glmnet)
```

```
install.packages("copula")
library(copula)
rm(list=ls())
X<-cbind(floor(rCopula(n = 100,copula =
normalCopula(param = .8,dim = 3))*3),
floor(rCopula(n = 100,copula =
normalCopula(param = .82,dim = 3))*3),
floor(rCopula(n = 100,copula =
normalCopula(param = .75,dim = 2))*3))
#o comando anterior cria um grupo de
#correlação 0.8 nas três primeiras
#covariaveis, 0.82 nas três seguintes
#e 0.75 nas duas últimas.
x1<-X[,1]
x2<-X[,2]
x3<-X[,3]
x4<-X[,4]
x5<-X[,5]
x6<-X[,6]
x7<-X[,7]
x8<-X[,8]
X<-cbind(x1,x2,x3,x4,x5,x6,x7,x8)
y=130*x1+120.1*x2+150*x3-15*x4+0.2*x5-0.15*x6
+60*x7-59*x8+rnorm(100,sd=0.1)
reg=glmnet(X,y,alpha=0.5)
plot(reg,label=TRUE)
```

```
cvreg=cv.glmnet(X,y,alpha=0.5)
coef(cvreg,s="lambda.min")
print(reg)
```

Código 5

```
rm(list=ls())
library(glmnet)
gendad=read.table("C:/Doutorado/DoutFazendo/Defesa
/GenGWS.txt",header = TRUE)
names(gendad)
C=gendad[,64:300]
colunas=c()
for(i in 1:237){
  colunas[i]=paste("c",i,sep="")
}
colnames(C)=colunas
y=gendad[,11]#a coluna 11 eh o pcarc
j=90
k=100
X=C[,j:k]
X<-X[-c(which(is.na(y))),]
y=y[-c(which(is.na(y)))]
X=as.matrix(X)
reg=glmnet(X,y,alpha=0.5)#nlambda=500
plot(reg,label=TRUE)
```

Código 6

```
rm(list=ls())
library(glmnet)
gendad=read.table("C:/Doutorado/DoutFazendo/Defesa
/GenGWS.txt",header = TRUE)
names(gendad)
C=gendad[,64:300]
colunas=c()
for(i in 1:237){
  colunas[i]=paste("c",i,sep="")
}
colnames(C)=colunas
y=gendad[,5]#a coluna 5 eh o pcarc
j=1
k=237#ultimo=237
X=C[,j:k]
X=as.matrix(X)
reg=glmnet(X,y,alpha=0.5)#nlambda=500
plot(reg,label=TRUE)
coef(reg,s=1.2)
colSums(coef(reg,s=1.2) != 0)#conta quantos coef
#são não nulos.
a=c(36,82,129,144,148,162,190,206,208,211,219,229)
#a linha anterior cria um vetor com
#os 12 marcadores mais relevantes.
X=cbind(C[,a])
```

```
X=as.matrix(X)
y=gendad[,5]#a coluna 5 eh o pcarc
reg=glmnet(X,y,alpha=0.5)#nlambda=500
plot(reg,label=TRUE)
```

Código 7

```
rm(list=ls())
library(glmnet)
gendad=read.table("C:/Doutorado/posdefesa/Defesa/
Codigos/GenGWS.txt",header = TRUE)
names(gendad)
C=gendad[,64:300]
colunas=c()
for(i in 1:237){
  colunas[i]=paste("c",i,sep="")
}
colnames(C)=colunas
y=gendad[,52]#a coluna 52 eh o pH45
j=1
k=237
X=C[,j:k]
X<-X[-c(which(is.na(y))),]
y=y[-c(which(is.na(y)))]
X=as.matrix(X)
reg=glmnet(X,y,alpha=0.5)
plot(reg,label=TRUE)
colSums(coef(reg,s=0.051) != 0)#conta quantos coef
```

```
#são não nulos.
coef(reg,s=0.051)
a=c(11,13,49,54,64,86,126,133,154,157,186,199,234)
C=gendad[,64:300]
colunas=c()
for(i in 1:237){
  colunas[i]=paste("c",i,sep="")
}
colnames(C)=colunas
y=gendad[,52]#a coluna 52 eh o pH45
X=cbind(C[,a])
X<-X[-c(which(is.na(y))),]
y=y[-c(which(is.na(y)))]
X=as.matrix(X)
reg=glmnet(X,y,alpha=0.5)
plot(reg,label=TRUE)
```

Código 8

```
rm(list=ls())
library(glmnet)
gendad=read.table("C:/Doutorado/posdefesa/Defesa/
Codigos/GenGWS.txt",header = TRUE)
names(gendad)
C=gendad[,64:300]
colunas=c()
for(i in 1:237){
  colunas[i]=paste("c",i,sep="")
```

```
}
colnames(C)=colunas
y=gendad[,52]
j=57
k=110
X=C[,j:k]
X<-X[-c(which(is.na(y))),]
y=y[-c(which(is.na(y)))]
X=as.matrix(X)
reg=glmnet(X,y,alpha=0.5)
plot(reg,label=TRUE)
colSums(coef(reg,s=0.031) != 0)#conta quantos coef
#são não nulos.
coef(reg,s=0.031)
a=c(64,67,69,72,76,81,86,87,101)
C=gendad[,64:300]
colunas=c()
for(i in 1:237){
  colunas[i]=paste("c",i,sep="")
}
colnames(C)=colunas
y=gendad[,52]
X=cbind(C[,a])
X<-X[-c(which(is.na(y))),]
y=y[-c(which(is.na(y)))]
X=as.matrix(X)
```

```
reg=glmnet(X,y,alpha=0.5)
plot(reg,label=TRUE)
```