Regularização em MLGs

Carlos de Moura

2025-12-01

Índice

Sobre

"Eu tô te explicando Pra te confundir Eu tô te confundindo Pra te esclarecer" Tom Zé

Este é o material auxiliar da apresentação do trabalho final do curso de Modelos Lineares Generalizados (MLGs) (DEST-UFMG, 2025/2). O tema é regularização em MLGs, em específico os métodos de regularização ridge, lasso e elastic net e este quarto book está dividido da seguinte forma:

- Definição de regularização e apresentação dos métodos de shrinkage, com exemplos de regressão normal;
- Estimação dos parâmetros em MLG com penalização;
- Tunagem dos parâmetros de regularização;
- Exemplo prático com dados reais no R.

Referências bibliográficas importantes que foram usadas para a feitura desse trabalho são citadas ao final do documento.

Alguns pacotes que usaremos estão abaixo listados.

```
if (!{"pak" %in% rownames(installed.packages())}) install.packages("pak")

pak::pak(c("matrixcalc", "glmnet"))
```

Motivação

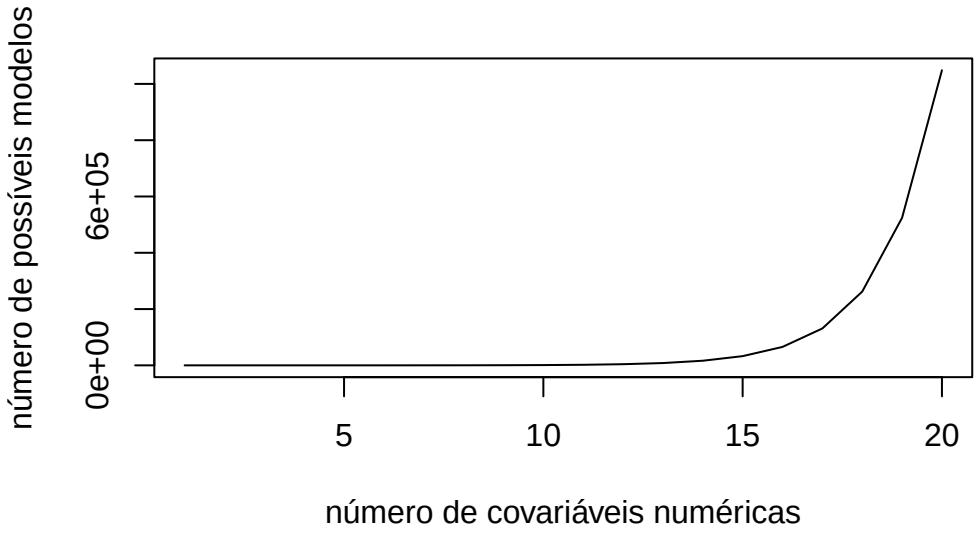
Porquê fazer seleção de modelos? Porquê regularizar coeficientes é uma boa ideia?

1 Regularização

1.1 Seleção de variáveis naïve

Stepwise

```
f = function(n) {
  s = 0
  for (i in 1:n) {
    s = s + choose(n, i)
  }
  s
}

seq(1, 20, 1) |>
  lapply(f) |>
  unlist() |>
  plot(type = 'l',
       main = "",
       xlab = "número de covariáveis numéricas",
       ylab = "número de possíveis modelos"
       )
```



1.2 Ridge

A técnica Ridge foi a primeira das três técnicas a surgir, no trabalho de Hoerl & Kennard (1970). Originalmente, os autores buscavam entender como superar problemas em que a matriz de covariáveis X estava mal-especificada. Relembrando que na regressão linear normal, temos que

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I_n)$$

O estimador de máxima verossimilhança (EMV) de β será o mesmo estimador de mínimos quadrados (EMQ)

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y,$$

se:

- $X^\top X$ for inversível;
- A matriz de covariáveis é ortogonalizável, i.e., os dados foram coletados de maneira independente;
- há menos betas que observações, isto é `ncol(X)` « `nrow(X)`.

Além disso, $\hat{\beta}$ é não viciado, consistente e tem matriz de covariâncias dada por

$$cov(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}.$$

Uma vez que temos a distribuição do estimador, fica fácil fazer inferência via intervalos de confiança, por exemplo.

$$\hat{\beta} \sim N_p\left(\beta,\ \sigma^2(X^\top X)^{-1}\right).$$

Nesse sentido, se temos uma matriz de dados problemática - no sentido em que $(X^\top X)^{-1}$ não está bem definida, teremos problema de estimação via EMQ.

Veja o exemplo numérico abaixo.

```
set.seed(12345)

n = 10
beta = c(1, 0)
X = cbind(1:n, 2*(1:n))
Y = X %*% beta + rnorm(n)
```

Ver matriz X

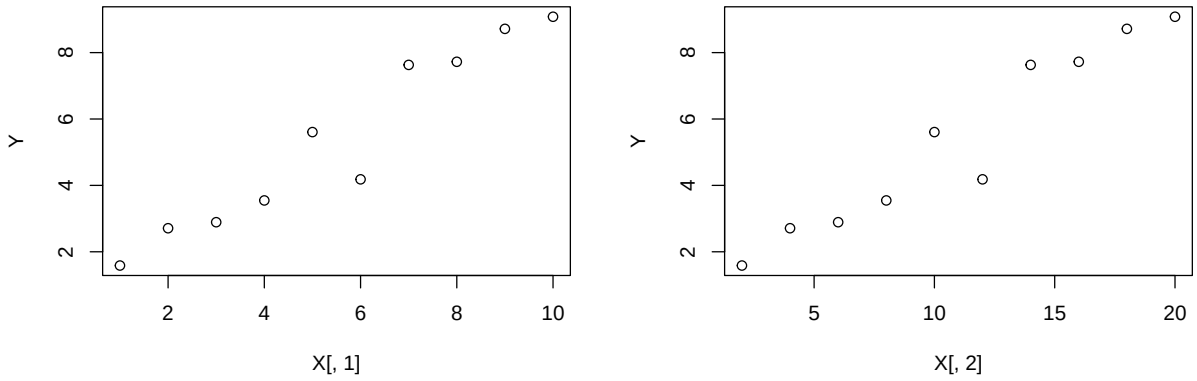
```
head(X)
```

```
     [,1] [,2]
[1,]    1    2
[2,]    2    4
[3,]    3    6
[4,]    4    8
[5,]    5   10
[6,]    6   12
```

```
matrixcalc::is.singular.matrix(t(X)%*%X)
```

```
[1] TRUE
```

```
par(mfrow=c(1,2))
plot(Y~X[,1])
plot(Y~X[,2])
```




```
lm(Y~0+X)
```

```
Call:
lm(formula = Y ~ 0 + X)

Coefficients:
    X1      X2
0.9544      NA
```

E se a regressão estivesse na outra covariável?

```
set.seed(12345)

beta = c(0, 1)
X = cbind(1:n, 2*(1:n))
Y = X %*% beta + rnorm(n)

lm(Y~0+X)
```

```
Call:
lm(formula = Y ~ 0 + X)

Coefficients:
   X1     X2
1.954     NA
```

Detalhes

Comentários sobre o modelo não saber selecionar variáveis.

```r
set.seed(12345)

beta = c(1, 0)
X = cbind(1:n, 2*(1:n))
ruido = cbind(rep(0,n), rnorm(n,0,.1))
X = X + ruido
Y = X %*% beta + rnorm(n)
```

Ver matriz X

```r
head(X)
```

```
     [,1]      [,2]
[1,]    1  2.058553
[2,]    2  4.070947
[3,]    3  5.989070
[4,]    4  7.954650
[5,]    5 10.060589
[6,]    6 11.818204
```

```r
matrixcalc::is.singular.matrix(t(X)%*%X)
```

```
[1] FALSE
```

```r
eigen(t(X)%*%X)$values
```

```
[1] 1.918025e+03 1.070613e-02
```

```r
lm(Y~0+X)
```

```

Call:
lm(formula = Y ~ 0 + X)

Coefficients:
    X1      X2
 6.658  -2.820
```

Comentários sobre combinações lineares, parcimônia e inflação da variância na inferência.

A proposta de Hoerl & Kennard (1970) é adicionar um certo múltiplo da matriz identidade à $X^\top X$, de modo que seja possível superar os problemas de uma matriz X mal especificada. James, Witten, Hastie, & Tibshirani (2021) mostra que isso é equivalente a adicionar uma punição no processo de estimação por mínimos quadrados. Seja $SQRes = (Y - X\beta)^\top (Y - X\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$, assim:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ SQRes + \lambda_{ridge} \sum_{j=1}^{p} \beta_j^2 \right\},$$

em que λ_{ridge} é uma constante positiva (escolhida pelo pesquisador) que dá o peso dessa punição.

Note que:

- À medida que λ cresce menores ficam as estimativas dos betas, $\hat{\beta}_{ridge} \to 0,\ \lambda_{ridge} \to \infty$.
- Não faz sentido incluir o intercepto no processo de shrinkage, pois $g(\beta_0)$ é o valor esperado da variável resposta dado que as demais covariáveis são zero, em que g é uma função de ligação.
- Fica claro que a depender do valor de λ que escolhemos, estamos incluindo algum viés no modelo.
- A estimação ridge diminui as estimativas dos betas na mesma proporção (se comparada com o MMQ tradicional).
- Em diminuindo a estimativa dos betas, diminui-se também a variância envolvida nas estimativas dos parâmetros.

> **Aviso**
>
> Ao usar métodos de regularização, estamos fazendo um trade-off entre variância e viés. Grosso modo, estamos inserido algum viés em nossas estimativas, a fim de diminuir a variância na estimação dos parâmetros e assim fazer inferências mais precisas.
> Um bom método de escolha do parâmetro de shrinkage é crucial para o bom funcionamento da estimação ridge. Falaremos disso nas próximas seções. *Spoiler: faremos validação cruzada.*

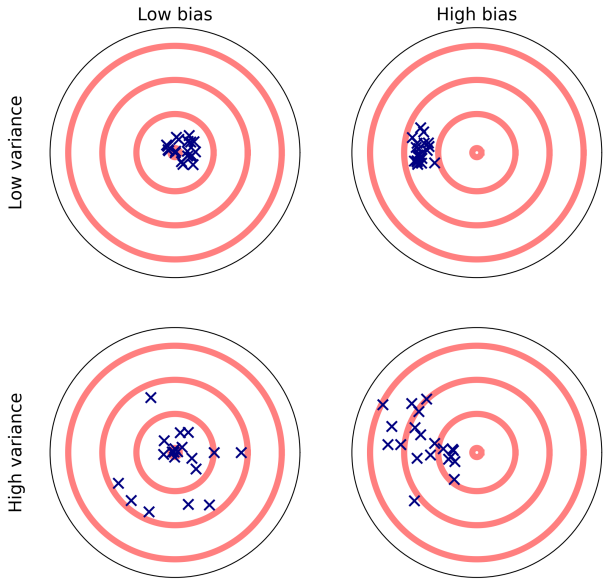


Figura 1.1: Analogia dos dardos para pensar variância e viés.

O exemplo abaixo está no livro James et al. (2021), `Credit` é um conjunto de dados incluído no pacote ISLR, contendo informações sobre 4000 clientes de um banco, com o objetivo de modelar e entender fatores associados ao limite de crédito de cada pessoa.

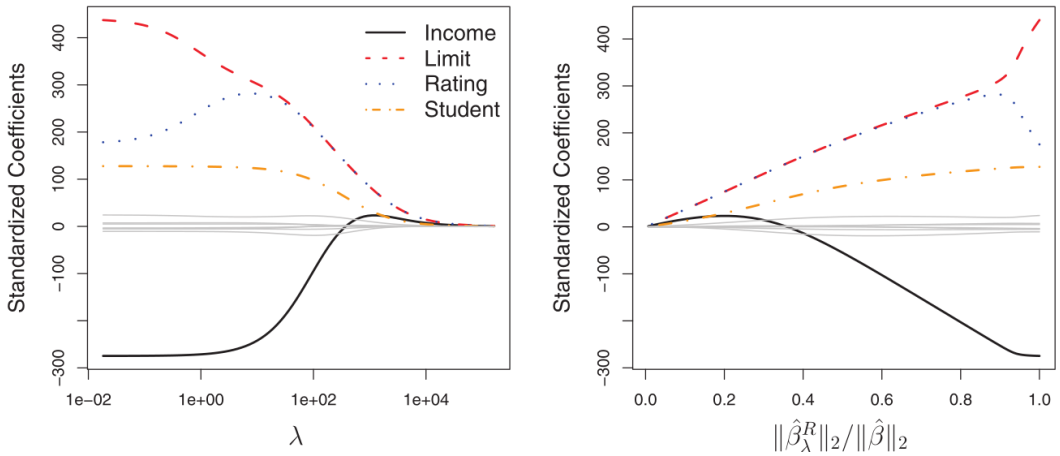


Figura 1.2: Exemplo de regressão ridge.

> **Padronização das covariáveis**
>
> A técnica ridge pode ser muito sensível à escala das covariáveis. Nesse sentido, vale a pena padronizar as covariáveis pelo desvio-padrão para evitar a influência desse efeito.

1.3 Lasso

A regressão lasso surgiu com o artigo de Tibshirani (1996). No caso da regressão linear normal, a diferença entre lasso e ridge pode ser vista na função de perda da estimação mínimos quadrados. Em vez de usar a norma ℓ_2 (como faz a regressão ridge), a regressão lasso usa a norma ℓ_1 de beta.

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ SQRes + \lambda_{lasso} \sum_{j=1}^{p} |\beta_j| \right\},$$

Muitas das observações que fizemos para a estimação ridge se aplicam tambpem à regressão lasso. A principal diferênça entre essas técnicas está no fato que a regressão lasso faz, de fato, uma seleção de variáveis. Isso se dá pois a regressão lasso pode zerar os coeficientes de algumas covariáveis.

> **Nota**
>
> O fato de a regressão lasso poder zerar os coeficientes das covariáveis está associado com a geometria imposta no espaço paramétrico.

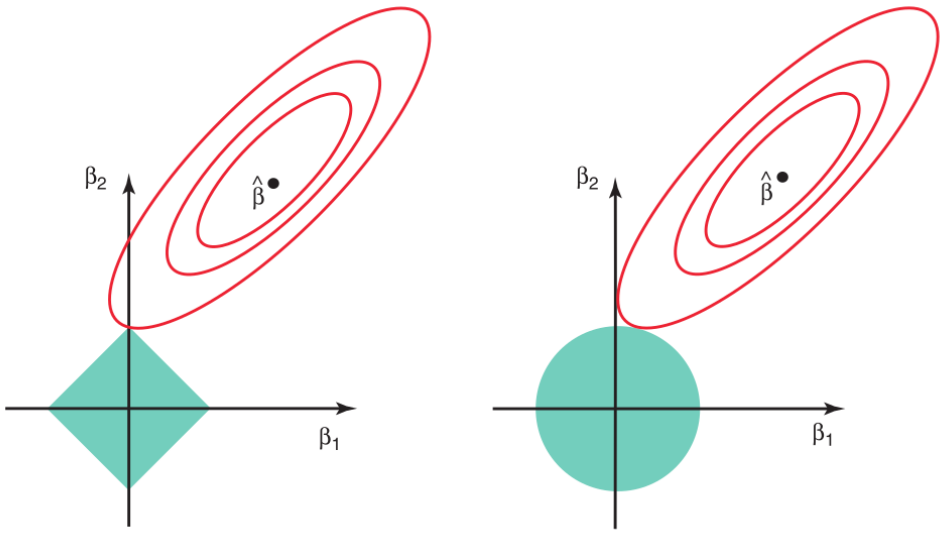


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

No mesmo banco de dados `Credit`, James et al. (2021) aplicou a regressão lasso:

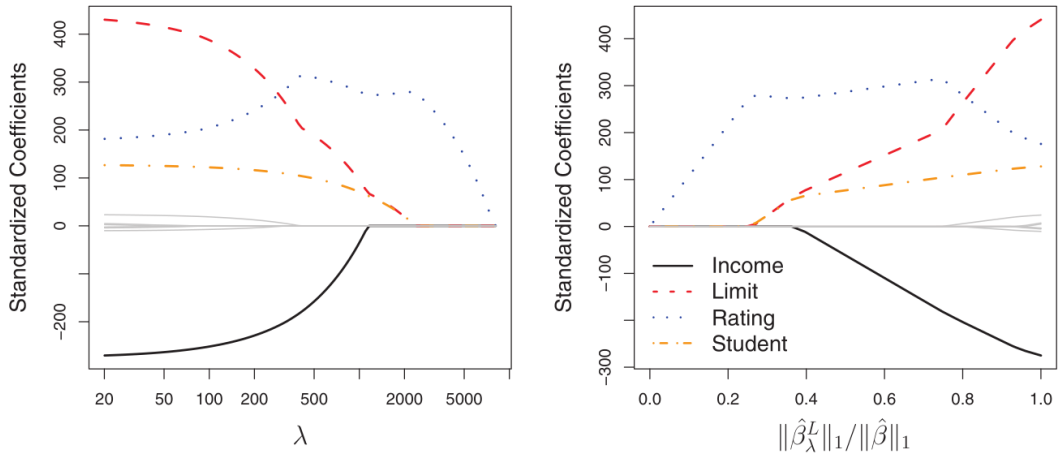


Figura 1.3: Exemplo de regressão lasso.

1.4 Comparação das técnicas

"You Can't Always Get What You Want"

A

1.4.1 Estudo sobre MSE com dados sintéticos

```
sqres = function(beta, X, Y) {
  res = Y - X %*% beta
  drop(crossprod(res))
}

sqres_ridge = function(beta, X, Y, lambda) {
  res = Y - X %*% beta
  drop(crossprod(res) + lambda * crossprod(beta))
}

sqres_lasso = function(beta, X, Y, lambda) {
  res = Y - X %*% beta
  drop(crossprod(res) + lambda * sum(abs(beta)))
}
```

A

> **Nota**
>
> Note que minimizar `sqres` deve dar o mesmo resultado que `lm`.

```
fit = optim(
  par    = rep(0,ncol(X)),
  fn     = sqres,
  X      = X,
  Y      = Y,
  method = "BFGS"
)

round(fit$par, 5) |> `names<-`(paste0("X", 1:ncol(X)))
```

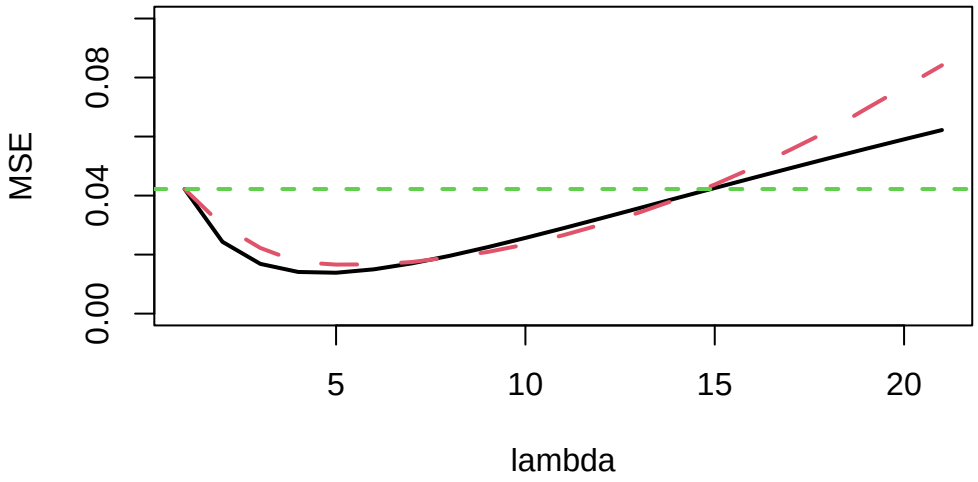
```
      X1       X2
 6.65825 -2.81988
```

```
lm(Y~0+X)
```

```
Call:
lm(formula = Y ~ 0 + X)

Coefficients:
    X1      X2
 6.658  -2.820
```

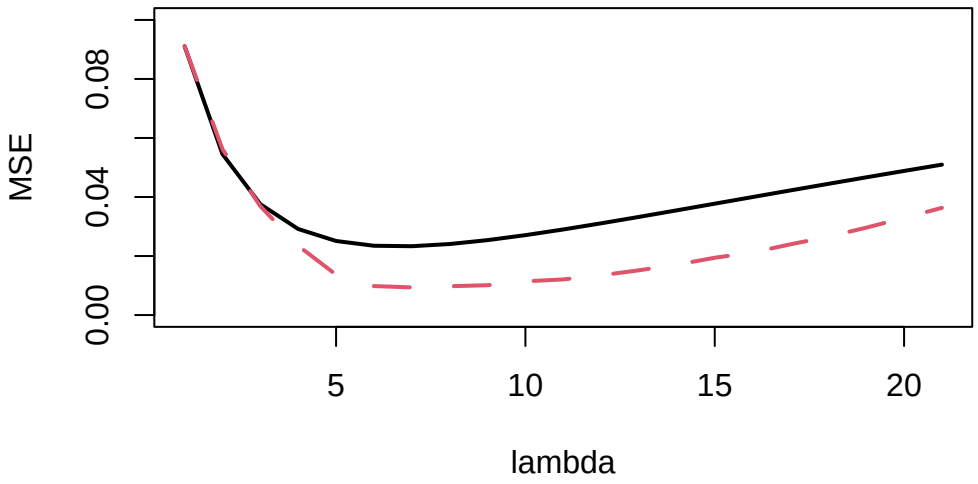
```r
set.seed(12345)
n = 20
beta = exp(-seq(-2,2,.5)^2)
X = rnorm(n*length(beta)) |> matrix(nrow = n, ncol = length(beta))
Y = X %*% beta + rnorm(n,0,1)

mse_ridge = mse_lasso = numeric()

for (lambda in 0:20) {
  beta_ridge =
    optim(
      par    = rep(0,ncol(X)),
      fn     = sqres_ridge,
      X      = X,
      Y      = Y,
      lambda = lambda,
      method = "BFGS"
    )$par

    mse_ridge[lambda + 1] = mean((beta_ridge - beta)^2)

  beta_lasso =
    optim(
      par    = rep(0,ncol(X)),
      fn     = sqres_lasso,
      X      = X,
      Y      = Y,
      lambda = lambda,
      method = "BFGS"
    )$par

    mse_lasso[lambda + 1] = mean((beta_lasso - beta)^2)
}

plot(mse_ridge,  type="l", lwd = 2, col = 1,
```

```
     ylim = c(0, 0.1),
     ylab = "MSE", xlab = "lambda")
lines(mse_lasso, type="l", lwd = 2, col = 2, lty = "aa")

mse_lm = mean((lm(Y~0+X)$coeff - beta)^2)
abline(h = mse_lm, lwd = 2, col = 3, lty = "44")
```




```
set.seed(12345)
n = 20
beta = c( rep(1,3), rep(0,10) )
X = rnorm(n*length(beta)) |> matrix(nrow = n, ncol = length(beta))
Y = X %*% beta + rnorm(n,0,1)

mse_ridge = mse_lasso = numeric()

for (lambda in 0:20) {
  beta_ridge =
    optim(
      par    = rep(0,ncol(X)),
      fn     = sqres_ridge,
      X      = X,
      Y      = Y,
      lambda = lambda,
      method = "BFGS"
    )$par

    mse_ridge[lambda + 1] = mean((beta_ridge - beta)^2)
```

```r
  beta_lasso =
    optim(
      par    = rep(0,ncol(X)),
      fn     = sqres_lasso,
      X      = X,
      Y      = Y,
      lambda = lambda,
      method = "BFGS"
    )$par

    mse_lasso[lambda + 1] = mean((beta_lasso - beta)^2)
}

plot(mse_ridge,  type="l", lwd = 2, col = 1,
     ylim = c(0, 0.1),
     ylab = "MSE", xlab = "lambda")
lines(mse_lasso, type="l", lwd = 2, col = 2, lty = "aa")
```



1.5 Elastic net

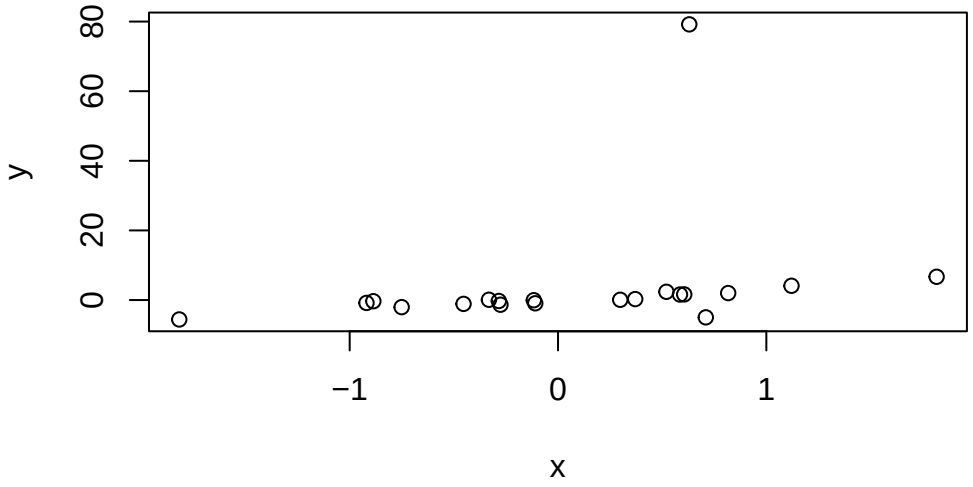
"Can you?"

O artigo Zou & Hastie (2005) introduz o elastic net, que nada mais é que uma mistura das técnicas ridge e lasso. A ideia inicial considera o seguinte estimador:

$$\begin{aligned}\hat{\beta}_{en}^{nave} &= \arg\min_{\beta} \left\{ SQRes + \lambda_{ridge} \sum_{j=1}^{p} \beta_j^2 + \lambda_{lasso} \sum_{j=1}^{p} |\beta_j| \right\} \\ &= \arg\min_{\beta} \left\{ SQRes + \lambda_{en} \left(\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \right) \right\},\end{aligned}$$

1.6 Lidando com outliers

```
set.seed(12345); x = rnorm(20); e = rt(20, 1); y = 2*x + e; plot(y~x); lm(y~0+x)
```




```
Call:
lm(formula = y ~ 0 + x)

Coefficients:
    x
6.136
```

2 Estimação

"Se isto for possível, Pois, me contem, Como escrever de novo, Um jornal de ontem" Tom Zé

2.1 Regularização como uma função de perda

2.2 Regularização como uma restrição do espaço paramétrico

É possível entender cada um dos processos de regularização descritos anteriormente como uma restrição do espaço paramétrico dos coeficientes de regressão. Se não fazer seleção é considerer que $\beta \in \mathbb{R}^d$, é possível mostrar que - escolhidos os parâmetros de shrinkage - então minimar a soma de quadrados do resíduo penalizada é a mesma coisa que minimizar a soma de quadrado da regressão num espaço paramétrico menor (que depende dos parâmetros de shrinkage escolhidos).

Assim, (na regressão linear normal) vale que:

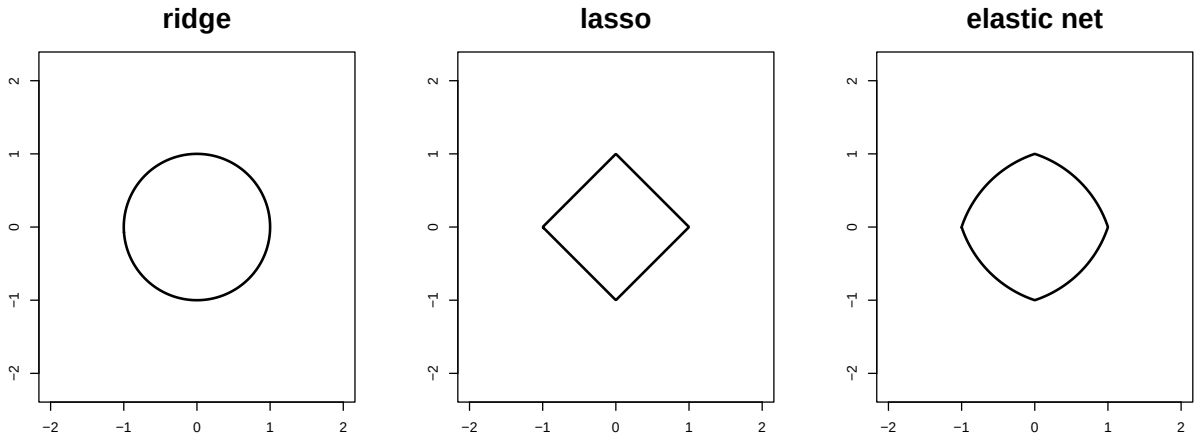
$$\hat{\beta}_{ridge} = \arg\min_{\beta}\{SQRes\} \quad \text{com } \beta \text{ tal que} \quad \sum_{j=1}^{p} \beta_j^2 \leq t_{ridge},$$

$$\hat{\beta}_{lasso} = \arg\min_{\beta}\{SQRes\} \quad \text{com } \beta \text{ tal que} \quad \sum_{j=1}^{p} |\beta_j| \leq t_{lasso},$$

$$\hat{\beta}_{elastic\ net} = \arg\min_{\beta}\{SQRes\} \quad \text{com } \beta \text{ tal que} \quad (1-\alpha)\sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t_{elastic\ net}.$$

```
desenhar_espaço_paramétrico = function(alpha, t, título = NULL) {
  stopifnot(
    "alpha deve estar entre 0 e 1" = all(alpha >= 0, alpha <= 1),
    "t deve ser positivo"          = t > 0
  )

  F = function(x, y) alpha*(x^2 + y^2 - t) + (1-alpha)*(abs(x) + abs(y) - t)
```

```
  x = seq(-2, 2, length = 400)
  y = seq(-2, 2, length = 400)
  g = outer(x, y, F)

  contour(x, y, g,
          levels = 0,
          drawlabels = FALSE,
          lwd = 2, asp = 1,
          main = título,
          cex.main = 2
          )
}


par(mfrow = c(1,3))
desenhar_espaço_paramétrico(1,   1, "ridge")
desenhar_espaço_paramétrico(0,   1, "lasso")
desenhar_espaço_paramétrico(1/2, 1, "elastic net")
```



2.3 Regularização nos MLGs

O modelo linear normal pode ser especificado diretamente por meio de seus resíduos, mas - em geral - isso não é possível em todos os MLGs. Aqui, apresentaremos a técnica descrita em Tay, Narasimhan, & Hastie (2023), que generaliza a regularização elastic net para os MLGs. Grosso modo, em vez de minimizar a SQRes penalizada (que nem sempre está bem definida), vamos minimizar o inverso aditivo da log-verossimilhança penalizada (que sempre está bem definida num MLG).

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0,\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \text{loglik}\,(y_i, X\beta) + P_{\alpha,\lambda}(\beta) \right\},$$

onde $P_{\alpha,\lambda}(\beta)$ é a função de penalização elastic net, ou seja

$$P_{\alpha,\lambda}(\beta) = \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right).$$

Com isso, temos um algoritmo de mínimos quadrados ponderados (IRLS) dado a seguir.

IRLS elastic net

Selecione um valor de $\alpha \in [0,1]$ e valor de $\lambda \in \mathbb{R}$.
Inicialize o algritmo com $(\hat{\beta}_0^{(0)}, \hat{\beta}^{(0)}) = (0,0)$. Doravante para $t = 1, 2, ...$ (até que se atinja convergência) faça:

1. Compute $\eta_i^{(t)} = \hat{\beta}_0^{(t)}(\lambda) + x_i^\top \hat{\beta}^{(t)}(\lambda)$ e $\mu_i^{(t)} = g^{-1}(\eta_i^{(t)})$, para $i \in 1:n$;
2. Compute $z_i^{(t)} = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \left(\frac{d\mu_i^{(t)}}{d\eta_i^{(t)}} \right)$ e também $w_i^{(t)} = \frac{1}{V(\mu_i^{(t)})} \left(\frac{d\mu_i^{(t)}}{d\eta_i^{(t)}} \right)^2$, para $i \in 1:n$;
3. Resolva

$$\left(\hat{\beta}_0^{(t+1)}(\lambda),\ \hat{\beta}^{(t+1)}(\lambda) \right) = \arg\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i^{(t)} (z_i^{(t)} - X\beta)^2 + P_{\alpha,\lambda}(\beta) \right].$$

2.4 Estudo Bootstrap

A

3 Tunagem

Validação cruzada

```
1
```

```
[1] 1
```

4 Exemplo prático

"Minha jangada vai sair por mar"

Esparsidade, multicolinearidade e outliers

```
1
```

```
[1] 1
```

Referências

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. Obtido de https://www.jstatsoft.org/index.php/jss/article/view/v033i01

Friedman, J., Hastie, T., Tibshirani, R., et al. (2025). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. Stanford University. Obtido de https://glmnet.stanford.edu/

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. Obtido de https://www.jstor.org/stable/1267351

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts em Statistics. Springer. Obtido de https://link.springer.com/book/10.1007/978-1-0716-1418-1

Kuhn, M., & Silge, J. (2022). *Tidy Modeling with R*. O'Reilly Media. Obtido de https://www.tmwr.org/

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Obtido de https://www.r-project.org/

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, *106*(1), 1–31. Obtido de https://www.jstatsoft.org/article/view/v106i01

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Obtido de https://www.jstor.org/stable/2346178

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320. Obtido de https://www.jstor.org/stable/3647580