Modelagem estatística em alta dimensão

Curso de Verão IMPA Tech – IDOR

Rodney Fonseca

Departamento de Estatística
Universidade Federal da Bahia

12 Jan 2026

Introdução

Apresentação

- **Nome:** Rodney Fonseca
- **Formação:** graduação (UFC), mestrado (UFPE) e doutorado (Unicamp) em Estatística
  Pós-doc no dept. de matemática aplicada e computação do Instituto Weizmann de Ciências em Israel
- **Pesquisa:** modelos de regressão, métodos não-paramétricos, dados de alta dimensão

Conteúdo das aulas

- Modelagem estatística em alta dimensão
  - Alta dimensionalidade e o conceito de esparsidade
  - Lasso para modelos lineares
  - Aplicação do lasso em modelos lineares generalizados (MLG)
- Inferência estatística com o estimador lasso
  - Lasso Bayesiano
  - Bootstrap
  - Inferência via lasso deviesado
- Material disponível em `rodneyfv.github.io`

Referências

- Livro-texto
  - Hastie, T., R. Tibshirani, and M. Wainwright (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. Boca Raton: CRC Press.
  - Disponível neste link: `https://hastie.su.domains/StatLearnSparsity/`

- Referências complementares
  - Izbicki, R. and T. M. dos Santos (2020). Aprendizado de máquina: uma abordagem estatística.
  - Fan, J., R. Li, C.-H. Zhang, and H. Zou (2020). Statistical Foundations of Data Science. Boca Raton: CRC press.
  - Wainwright, M. J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press.

O problema de alta dimensionalidade e o conceito de esparsidade

Alguns motivos para coleta massiva de dados

- Descobrir indicadores para desenvolvimento de doenças através de transcrições genéticas
- Identificar biomarcadores para o tempo de vida de pacientes
- Encontrar conexões entre regiões cerebrais associadas com certa característica

“Estamos nos afogando em informação e famintos por conhecimento” (Hastie et al., 2015, p. 1)

- Necessidade de extrair informações úteis de grandes volumes de dados

"Estamos nos afogando em informação e famintos por conhecimento" (Hastie et al., 2015, p. 1)

- Necessidade de extrair informações úteis de grandes volumes de dados
- **Ideia:** Supor que o mundo não é tão complexo quanto parece
  - Dentre milhares de genes, alguns estão diretamente ligados à doença
  - Alguns fatores afetam mais o tempo de vida (idade, peso, fuma?, etc.)
  - Algumas regiões cerebrais se ativam mais durante atividades específicas

- O número de preditores disponível pode ser grande, mas somente alguns são relevantes

Esparsidade

Suposição de que o fenômeno de interesse pode ser bem explicado por um modelo estatístico com um número pequeno de preditores

Regressão linear

- Considere que temos n observações de uma variável resposta y_i e p preditores associados $x_i = (x_{i1}, \ldots, x_{ip})^\top$
- O modelo de regressão linear é dado por

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \quad i = 1, \ldots, n,$$

em que β_0 e $\beta = (\beta_1, \ldots, \beta_p)^\top$ são parâmetros desconhecidos e e_i é um erro aleatório

- Estimativas por **mínimos quadrados** (MQ)

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

- Tipicamente, todas as estimativas são diferentes diferentes de zero, ou seja, todos os preditores tem alguma importância

- Estimativas por **mínimos quadrados** (MQ)

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

- Tipicamente, todas as estimativas são diferentes diferentes de zero, ou seja, todos os preditores tem alguma importância
- Problemas quando $p > n$
  - tal interpretação é mais difícil
  - soluções de MQ não são únicas
  - Sobreajuste (overfitting)

Alternativa: regularização/penalização

- Estimativas via **lasso** ou com regularização ℓ_1:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \quad \text{sujeito a} \quad \|\beta\|_1 \leq t,$$

  em que $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ é a norma ℓ_1 de β e $t \geq 0$ é um parâmetro de ajuste (tuning parameter)

- O parâmetro t controla a quantidade de preditores que parecem relevantes

Por que o lasso é especial?

- lasso pode gerar estimativas esparsas, ou seja, somente alguns β_j estimados diferentes de zero
  - Isso não ocorre com penalizações ℓ_q para $q > 1$, como penalização **ridge** ($q = 2$)
- lasso é obtido de um problema convexo, o que simplifica a otimização da função objetivo
  - Isso não ocorre com penalizações ℓ_q para $q < 1$, como **escolher o melhor subconjunto de preditores** ($q = 0$)

Aposta na esparsidade

- Se $p \gg n$ e o modelo verdadeiro não é esparso, então a amostra não é suficiente para estimar bem os parâmetros desconhecidos
  - Modelo não é identificável (o problema assume inúmeras soluções)

Aposta na esparsidade

- Se $p \gg n$ e o modelo verdadeiro não é esparso, então a amostra não é suficiente para estimar bem os parâmetros desconhecidos
  - Modelo não é identificável (o problema assume inúmeras soluções)
- Porém, se o modelo verdadeiro é esparso, com somente $k < n$ parâmetros diferentes de zero, então há uma chance destes parâmetros serem estimados
  - Métodos esparsos, como o lasso, são capazes de estimar tais modelos mesmo sem sabermos exatamente quais são o k parâmetros importantes

Exemplo

Figura: Dado em formato de uma imagem 512x512

- A imagem foi vetorizada e foi aplicada uma transformada de wavelet
- São os mesmos dados, só que numa base diferente
- A imagem ao lado é uma amostra de 5.000 dos $512^2 = 262.144$ coeficientes

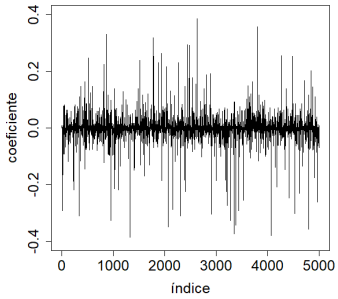

Figura: Coeficientes de wavelet

- Será que esse dado admite uma representação esparsa?
- Para testar, os 80% menores coeficientes (em valor absoluto) foram encolhidos para ter valor zero

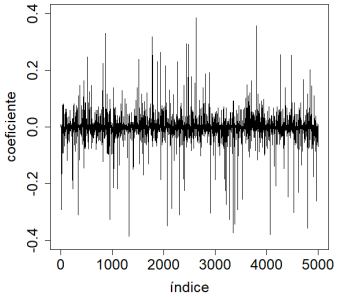

Figura: Coeficientes de wavelet

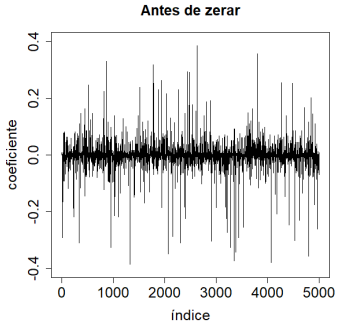

Figura: Coeficientes de wavelet (975 zeros)

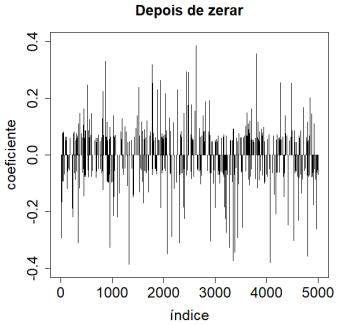

Figura: Versão esparsa desses coeficientes (4492 zeros)

Antes e depois das imagens

Figura: Versão original

Figura: Versão esparsa

O estimador lasso

Regressão linear

- O estimador de mínimos quadrados (MQ) é amplamente utilizado e tem boas propriedades[1]:
  - É não viesado
  - É o estimador linear não viesado de variância mínima (BLUE)
  - Tem distribuição assintótica normal

[1] Sob certas suposições no modelo

Regressão linear

- O estimador de mínimos quadrados (MQ) é amplamente utilizado e tem boas propriedades[1]:
  - É não viesado
  - É o estimador linear não viesado de variância mínima (BLUE)
  - Tem distribuição assintótica normal
- Alguns motivos para considerar alternativas:
  - Melhor generalização/previsão: permitindo algum viés na estimativas dos parâmetros, possivelmente podemos achar estimadores com menor variabilidade em termos de previsão.
  - Maior interpretabilidade: com poucas estimativas diferentes de zero, é mais simples explicar quais preditores são mais relevantes num modelo.

[1] Sob certas suposições no modelo

- O lasso (*least absolute selection and shrinkage operator*) foi proposto por Tibshirani (1996)

Figura: `https://hastie.su.domains/StatLearnSparsity/images/jsm_booksigning_2015.jpg`

- Representaremos pares de resposta-preditor $\{(y_i, x_i)\}_{i=1}^n$ com um vetor $\mathbf{y} = (y_1, \ldots, y_n)$ de respostas e uma matriz $\mathbf{X} \in \mathbb{R}^{n \times p}$ cuja i-ésima linha contém a observação $x_i^\top \in \mathbb{R}^p$

- O estimador lasso acha a solução $(\hat{\beta}_0, \hat{\beta})$ da minimização com restrição

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_2^2 \right\} \quad \text{suj.} \quad \|\beta\|_1 \leq t,$$

em que $\mathbf{1}$ é um vetor de n uns, $\|\cdot\|_1$ é a norma ℓ_1 e $\|\cdot\|_2$ é a norma Euclidiana

- Tipicamente as colunas de **X** são **padronizadas** para terem média zero ($\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$) e variância um ($\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$)
  - Evita estimativas serem afetadas pela unidade de medida

- Tipicamente as colunas de **X** são **padronizadas** para terem média zero ($\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$) e variância um ($\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$)
  - Evita estimativas serem afetadas pela unidade de medida
- A resposta também é centralizada para ter média zero ($\frac{1}{n}\sum_{i=1}^{n} y_i = 0$)
  - Podemos **omitir o intercepto** β_0 da otimização
  - O intercepto pode ser estimado como $\hat{\beta}_0 = \bar{y}$, a média original da variável resposta

- Uma versão alternativa da função objetivo do lasso é a sua **forma Lagrangiana**

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

em que $\lambda \geq 0$ é uma penalização

- As formas Lagrangiana e de minimização restrita são equivalentes, i.e., para cada λ existe um t que gera a mesma solução

- Valores de λ tipicamente são escolhidos via validação cruzada

- Uma alternativa ao lasso é a **regressão ridge**
- Método que pode ser representado como MQ penalizados com a norma ℓ_2:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{2} \|\beta\|_2^2 \right\},$$

em que $\lambda \geq 0$ é uma penalização

- Uma alternativa ao lasso é a **regressão ridge**
- Método que pode ser representado como MQ penalizados com a norma ℓ_2:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{2} \|\beta\|_2^2 \right\},$$

  em que $\lambda \geq 0$ é uma penalização
- Vantagens: problema convexo e função objetivo é diferenciável

- Uma alternativa ao lasso é a **regressão ridge**
- Método que pode ser representado como MQ penalizados com a norma ℓ_2:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{2} \|\beta\|_2^2 \right\},$$

  em que $\lambda \geq 0$ é uma penalização
- Vantagens: problema convexo e função objetivo é diferenciável
- Desvantagem: não faz seleção de variáveis[2]

[2] é possível usando *thresholding* (Zhang and Politis, 2022)

- O valor da penalização λ controla a complexidade do modelo
  - λ baixo: melhor ajuste, mas risco de sobreajuste (overfitting)
  - λ alto: ajuste esparso e mais interpretável, mas risco de "perder o sinal"

- O valor da penalização λ controla a complexidade do modelo
  - λ baixo: melhor ajuste, mas risco de sobreajuste (overfitting)
  - λ alto: ajuste esparso e mais interpretável, mas risco de "perder o sinal"
- Intuito é achar λ com um bom balanço entre esses cenários, algo geralmente feito via **validação cruzada** (cross-validation)
  - Definimos um grid de λ's e estimamos o erro quadrático médio ($\widehat{EQM}$) para cada um deles
  - Escolhemos λ que produz o menor $\widehat{EQM}$

Introdução ao MLG

- Até o momento, focamos no modelo linear ajustado com mínimos quadrados (penalizados)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Vantagens
  - Interpretação simples
  - Muito útil quando ϵ tem distribuição normal
  - Diversas formas de otimização possíveis

- Até o momento, focamos no modelo linear ajustado com mínimos quadrados (penalizados)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Vantagens
  - Interpretação simples
  - Muito útil quando ϵ tem distribuição normal
  - Diversas formas de otimização possíveis

- Desvantagem: pode ser altamente inadequado para certos tipos de variáveis resposta

Quando o modelo linear simples normal é inadequado

- Dados binários indicando presença ou ausência de algum atributo (doente vs. saudável). Dist. binomial é mais adequada

Quando o modelo linear simples normal é inadequado

- Dados binários indicando presença ou ausência de algum atributo (doente vs. saudável). Dist. binomial é mais adequada
- Dados de contagem (número de clientes que visitam uma loja num dia). Dist. Poisson é mais adequada

Quando o modelo linear simples normal é inadequado

- Dados binários indicando presença ou ausência de algum atributo (doente vs. saudável). Dist. binomial é mais adequada
- Dados de contagem (número de clientes que visitam uma loja num dia). Dist. Poisson é mais adequada
- Dados positivos (tempo de funcionamento de um aparelho eletrônico). Dist. gama é mais adequada

Exemplo: variáveis binárias

- A resposta é $Y \in \{0, 1\}$ (eg., 1 se doente e 0 se saudável)
- Frequentemente analisada com o modelo logístico

$$\log\left(\frac{P(Y = 1|X = \mathbf{x})}{P(Y = 0|X = \mathbf{x})}\right) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x},$$

em que $\mathbf{x} = (x_1, \ldots, x_p)$ é um vetor de preditores, β_0 e $\boldsymbol{\beta}$ são coeficientes de regressão

Exemplo: variáveis binárias

- O modelo logístico pode ser reescrito como

$$E(Y|X = x) = P(Y = 1|X = \mathbf{x}) = \frac{e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}}}{1 + e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}}}$$

- A média condicional é ligada ao preditor linear através da função logit $g(\mu) = e^{\mu}/(1 + e^{\mu})$

Exemplo: variáveis binárias

- O modelo logístico pode ser reescrito como

$$E(Y|X = x) = P(Y = 1|X = \mathbf{x}) = \frac{e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}}}{1 + e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}}}$$

- A média condicional é ligada ao preditor linear através da função logit $g(\mu) = e^{\mu}/(1 + e^{\mu})$
- Garante que valores ajustados estarão em $[0, 1]$

Exemplo: dados de contagem

- Resposta $Y \in \{0, 1, 2, \ldots\}$
- Frequentemente analisado com um modelo log-linear

$$\log E(Y|X = x) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

Exemplo: dados de contagem

- Resposta $Y \in \{0, 1, 2, \ldots\}$
- Frequentemente analisado com um modelo log-linear

$$\log E(Y|X = x) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

- $E(Y|X = x) = e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}}$ sempre fornece valores ajustados positivos

Modelos lineares generalizados

- Classe de modelos para variáveis respostas cuja distribuição pertence à família exponencial
  - Inclui normal, Bernoulli, binomial, Poisson, gama, etc.

- Classe de modelos para variáveis respostas cuja distribuição pertence à família exponencial
  - Inclui normal, Bernoulli, binomial, Poisson, gama, etc.
- Modelo de regressão direto na distribuição da resposta
- Média condicional $\mu(\mathbf{x}) = E(Y|X = \mathbf{x})$ ligada a um **preditor linear** por

$$g\Big(\mu(\mathbf{x})\Big) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x},$$

em que $g : \mathbb{R} \to \mathbb{R}$ é uma **função de ligação** monótona

Exemplos

- Resposta normal:

$$\mu(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} \quad \text{e} \quad g(\mu) = \mu$$

- Resposta Bernoulli:

$$\mu(\mathbf{x}) = P(Y = 1 | X = \mathbf{x}) \quad \text{e} \quad g(\mu) = \log\left(\mu/(1-\mu)\right)$$

- Resposta Poisson:

$$\mu(\mathbf{x}) = e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}} \quad \text{e} \quad g(\mu) = \log(\mu)$$

Estimação

- Feita maximizando a função de log-verossimilhança, digamos $\mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$

Estimação

- Feita maximizando a função de log-verossimilhança, digamos $\mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$
- Equivalente à calcular

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\beta_0, \boldsymbol{\beta}} \left\{ -\frac{1}{n} \mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) \right\}$$

- No caso de respostas Gaussianas, vira o problema de mínimos quadrados

Estimação com regularização

- Veremos versões regularizadas desses estimadores que promovem soluções esparsas

Estimação com regularização

- Veremos versões regularizadas desses estimadores que promovem soluções esparsas
- Estimativas calculadas minimizando o negativo da log-verossimilhança mais uma penalização ℓ_1:

$$\min_{\beta_0,\boldsymbol{\beta}} \left\{ -\frac{1}{n}\mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{y}, \mathbf{X}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

- Intercepto β_0 tipicamente não é penalizado

Medida de qualidade do ajuste

- Modelo linear normal: soma de quadrados dos resíduos $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$, em que $\hat{\mu}_i$ é o valor ajustado de y_i

Medida de qualidade do ajuste

- Modelo linear normal: soma de quadrados dos resíduos $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$, em que $\hat{\mu}_i$ é o valor ajustado de y_i
- Em MLG, tipicamente usamos a log-verossimilhança como $\mathcal{L}(\boldsymbol{\mu}, \mathbf{y})$, em que $\mu_i = g^{-1}\left(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}\right)$

Medida de qualidade do ajuste

- Modelo linear normal: soma de quadrados dos resíduos $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$, em que $\hat{\mu}_i$ é o valor ajustado de y_i
- Em MLG, tipicamente usamos a log-verossimilhança como $\mathcal{L}(\boldsymbol{\mu}, \mathbf{y})$, em que $\mu_i = g^{-1}\left(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}\right)$
- Qualidade medida pela **função desvio** (deviance)

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\Big\{ \underbrace{\mathcal{L}(\mathbf{y}, \mathbf{y})}_{\substack{\text{modelo saturado} \\ \hat{\mu}_i = y_i}} - \underbrace{\mathcal{L}(\hat{\boldsymbol{\mu}}, \mathbf{y})}_{\substack{\text{modelo ajustado} \\ \hat{\mu}_i = g^{-1}\left(\hat{\beta}_0 + \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\right)}} \Big\}$$

Regressão logística

- Modelo muito popular para variáveis binárias (problemas de classificação com duas classes)

- Modelo muito popular para variáveis binárias (problemas de classificação com duas classes)
- Em alta dimensão, regularização é necessária para estabilizar a otimização e interpretabilidade

- Modelo muito popular para variáveis binárias (problemas de classificação com duas classes)
- Em alta dimensão, regularização é necessária para estabilizar a otimização e interpretabilidade
- Exemplos práticos com p grande:
  - dados genéticos
  - imagens
  - palavras/tokens

- O objetivo é estimar a probabilidade condicional

$$P(Y = 1|X = \mathbf{x}) = E(Y|X = \mathbf{x})$$

- O objetivo é estimar a probabilidade condicional

$$P(Y = 1|X = \mathbf{x}) = E(Y|X = \mathbf{x})$$

- Usando a função de ligação logit, a função objetivo será

$$\begin{aligned} Q(\beta_0, \boldsymbol{\beta}) &= -\frac{1}{n}\sum_{i=1}^{n} \log\left(P(Y=1|X=\mathbf{x}_i)^{y_i} P(Y=0|X=\mathbf{x}_i)^{1-y_i}\right) + \lambda\|\boldsymbol{\beta}\|_1 \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log\left(y_i \cdot (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) - \log\left(1 + e^{\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}}\right)\right) + \lambda\|\boldsymbol{\beta}\|_1 \end{aligned}$$

- Deviance no modelo logístico:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \left\{ y_i \cdot g(\hat{\mu}_i) + \log\left(1 + e^{g(\hat{\mu}_i)}\right) \right\},$$

em que $g(\hat{\mu}_i) = \log\left(\hat{\mu}_i/(1-\hat{\mu}_i)\right) = \hat{\beta}_0 + \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$

- Deviance no modelo logístico:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \left\{ y_i \cdot g(\hat{\mu}_i) + \log\left(1 + e^{g(\hat{\mu}_i)}\right) \right\},$$

em que $g(\hat{\mu}_i) = \log\left(\hat{\mu}_i/(1-\hat{\mu}_i)\right) = \hat{\beta}_0 + \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$

- `glmnet` no R utiliza também a fração de deviance explicada em comparação com o modelo nulo ($\mu_i = \bar{y}$ para todo i)

$$D_\lambda^2 = 1 - \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}}_\lambda)}{D(\mathbf{y}, \bar{y} \cdot \mathbf{1})}$$

- Deviance pode ser utilizada para escolher λ

- Modelo logístico com ligação logit tem **razão de chances**

$$\frac{P(Y=1|X=\mathbf{x}_i)}{P(Y=0|X=\mathbf{x}_i)} = \frac{\mu_i}{1-\mu_i} = e^{\beta_0+\mathbf{x}_i^\top \boldsymbol{\beta}}$$

- Modelo logístico com ligação logit tem **razão de chances**

$$\frac{P(Y=1|X=\mathbf{x}_i)}{P(Y=0|X=\mathbf{x}_i)} = \frac{\mu_i}{1-\mu_i} = e^{\beta_0+\mathbf{x}_i^\top \boldsymbol{\beta}}$$

- Se x_j aumenta uma unidade e as demais covariáveis ficam fixas, i.e., $\tilde{x}_j = x_j + 1$ e $\tilde{x}_k = x_k$ para $k \neq j$, temos

$$\frac{P(Y=1|X=\tilde{\mathbf{x}}_i)}{P(Y=0|X=\tilde{\mathbf{x}}_i)} = \frac{\mu_i}{1-\mu_i} = e^{\beta_0+\tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}} = e^{\beta_0+\mathbf{x}_i^\top \boldsymbol{\beta}} \cdot e^{\beta_j}$$

- Modelo logístico com ligação logit tem **razão de chances**

$$\frac{P(Y=1|X=\mathbf{x}_i)}{P(Y=0|X=\mathbf{x}_i)} = \frac{\mu_i}{1-\mu_i} = e^{\beta_0+\mathbf{x}_i^\top \boldsymbol{\beta}}$$

- Se x_j aumenta uma unidade e as demais covariáveis ficam fixas, i.e., $\tilde{x}_j = x_j + 1$ e $\tilde{x}_k = x_k$ para $k \neq j$, temos

$$\frac{P(Y=1|X=\tilde{\mathbf{x}}_i)}{P(Y=0|X=\tilde{\mathbf{x}}_i)} = \frac{\mu_i}{1-\mu_i} = e^{\beta_0+\tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}} = e^{\beta_0+\mathbf{x}_i^\top \boldsymbol{\beta}} \cdot e^{\beta_j}$$

- Ex.: $\hat{\beta}_1 = 0{,}1 \Rightarrow e^{\hat{\beta}_1} \approx 1{,}10$ (aumento de 10% na chance de $Y = 1$)

Observações

- O intercepto β_0 não é penalizado no modelo logístico
  - $\lambda \to \infty$ leva ao modelo nulo com $\hat{\mu}_i = \bar{y}$

Observações

- O intercepto β_0 não é penalizado no modelo logístico
  - $\lambda \to \infty$ leva ao modelo nulo com $\hat{\mu}_i = \bar{y}$
- Preditores são padronizados para a penalização fazer sentido (padrão no `glmnet`)

Observações

- O intercepto β_0 não é penalizado no modelo logístico
  - $\lambda \to \infty$ leva ao modelo nulo com $\hat{\mu}_i = \bar{y}$
- Preditores são padronizados para a penalização fazer sentido (padrão no `glmnet`)
- Ajuste no `glmnet` é feito com as opções `family='binomial'` e `alpha=1`

Exemplo: dados de mamografia

- Classificar de pacientes com câncer de mama em casos benigno ou maligno (Fan et al., 2020, p. 254)
- Atributos oriundos de exames de imagens
- $n = 815$ observações e seis covariáveis, das quais duas são categóricas

- X_1: BI-RADS, um escore atribuído por médicos
- X_2: idade do paciente
- X_3: densidade de massa, 1 (alta) até 4 (baixa)
- X_4: dummy, 1 se forma circular e 0 c.c.
- X_5: dummy, 1 se forma oval e 0 c.c.
- X_6: dummy, 1 se forma lobular e 0 c.c.
  se $X_4 = X_5 = X_6 = 0$, a forma é irregular
- X_7: dummy, 1 se margem é circunscrita e 0 c.c.
- X_8: dummy, 1 se margem é microlobulada e 0 c.c.
- X_9: dummy, 1 se margem é obscura e 0 c.c.
- X_{10}: dummy, 1 se margem é má definida e 0 c.c.
  se $X_7 = X_8 = X_9 = X_{10} = 0$, a margem é espiculada

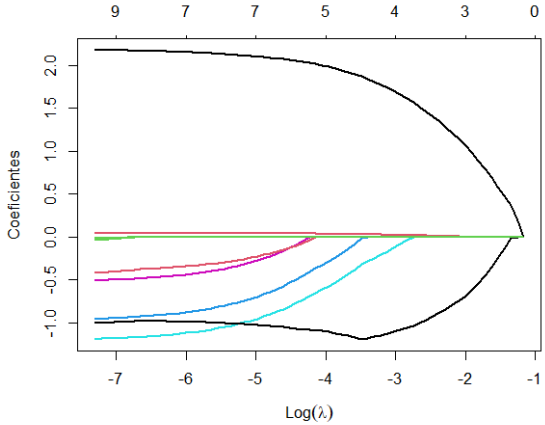

Figura: Trajetória das estimativas do lasso para diferentes penalizações λ

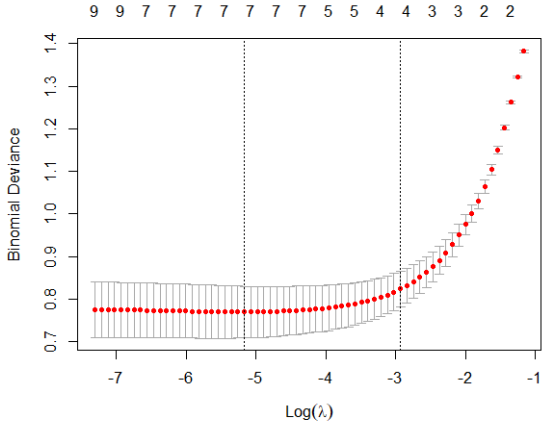

Figura: Deviance média estimada via validação cruzada para diferentes λ's. Mínimo atingido em 0,0057 (log -5,1647)

Tabela: Estimativas pelo lasso com $\lambda = 0,0057$ e pelo MLG usual

preditor	$\hat{\beta}$ (lasso)	$e^{\hat{\beta}}$ (lasso)	$\hat{\beta}$ (MLG)	$e^{\hat{\beta}}$ (MLG)
intercepto	-10.82	0	-0.95	0.39
birads	2.12	8.31	0.31	1.37
idade	0.04	1.04	0.01	1.01
densidade	0	1	-0.01	0.99
circular	-0.75	0.47	-0.19	0.82
oval	-1	0.37	-0.23	0.79
lobular	-0.32	0.73	-0.11	0.89
circunscrito	-1.02	0.36	-0.2	0.82
microlobulado	0	1	-0.01	0.99
obscuro	-0.26	0.77	-0.09	0.91
maldef	0	1	-0.02	0.98

Cuidado na interpretação

- Interpretação usual da razão de chances não vale se não levarmos em conta a **padronização dos preditores**

- Seja $x_j^p = x_j / s_j$ o j-ésimo preditor padronizado com fator de escala s_j. Aumentando uma unidade de x_j obtemos

$$\tilde{x}_j^p = \frac{x_j + 1}{s_j} = x_j^p + \frac{1}{s_j}$$

e a razão de chances vira

$$\frac{P(Y=1|X=\tilde{\mathbf{x}}_i^p)}{P(Y=0|X=\tilde{\mathbf{x}}_i^p)} = e^{\beta_0 + (\mathbf{x}_i^p)^\top \boldsymbol{\beta}} \cdot e^{\beta_j / s_j}$$

Modelo log-linear e MLG Poisson

- Poisson é bastante utilizada para modelar dados de contagem
- **Modelo log-linear** para a média:

$$Y|X = \mathbf{x} \sim Pois\big(\mu(\mathbf{x})\big), \quad \log \mu(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

- Poisson é bastante utilizada para modelar dados de contagem
- **Modelo log-linear** para a média:

$$Y|X = \mathbf{x} \sim Pois\big(\mu(\mathbf{x})\big), \quad \log \mu(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

- Garante previsões positivas $\hat{\mu}(\mathbf{x}) = e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^\top \mathbf{x}}$

Estimação

- Estimação feita minimizando o negativo da log-verossimilhança Poisson penalizada

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\beta_0+\boldsymbol{\beta}^\top\mathbf{x}_i)-e^{\beta_0+\boldsymbol{\beta}^\top\mathbf{x}}\right\}+\lambda\|\boldsymbol{\beta}\|_1$$

Estimação

- Estimação feita minimizando o negativo da log-verossimilhança Poisson penalizada

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\beta_0+\boldsymbol{\beta}^\top\mathbf{x}_i)-e^{\beta_0+\boldsymbol{\beta}^\top\mathbf{x}}\right\}+\lambda\|\boldsymbol{\beta}\|_1$$

- O intercepto β_0 tipicamente não é penalizado
  - Ajuste tende então ao modelo nulo $\hat{\mu}(\mathbf{x}_i)=\bar{y}$ conforme $\lambda\to\infty$

Qualidade do ajuste

- Função deviance no modelo Poisson:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{y_i}{\mu_i} \right) - (y_i - \hat{\mu}_i) \right\},$$

tomando o i-ésimo termo da soma como $\hat{\mu}_i$ caso $y_i = 0$

Qualidade do ajuste

- Função deviance no modelo Poisson:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{y_i}{\mu_i} \right) - (y_i - \hat{\mu}_i) \right\},$$

  tomando o i-ésimo termo da soma como $\hat{\mu}_i$ caso $y_i = 0$

- Em MLG usual, a deviance pode ser utilizada para fazer inferência por testes de hipóteses (Paula, 2025, p. 456)

Qualidade do ajuste

- Função deviance no modelo Poisson:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{y_i}{\mu_i} \right) - (y_i - \hat{\mu}_i) \right\},$$

  tomando o i-ésimo termo da soma como $\hat{\mu}_i$ caso $y_i = 0$

- Em MLG usual, a deviance pode ser utilizada para fazer inferência por testes de hipóteses (Paula, 2025, p. 456)
  - **Não vale** para o MLG Poisson com lasso
  - **Alternativas:** bootstrap ou lasso deviesado (van de Geer et al., 2014)

Interpretação

- Se o preditor x_j aumentar uma unidade e os demais ficarem fixos ($\tilde{x}_j = x_j + 1$ e $\tilde{x}_k = x_k$ para $k \neq j$), a média atualizada será

$$E(Y|X = \tilde{x}) = e^{\beta_0 + \boldsymbol{\beta}^\top \tilde{\mathbf{x}}} = e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}} \cdot e^{\beta_j}$$

- Fator de mudança de e^{β_j} no valor esperador

Interpretação

- Se o preditor x_j aumentar uma unidade e os demais ficarem fixos ($\tilde{x}_j = x_j + 1$ e $\tilde{x}_k = x_k$ para $k \neq j$), a média atualizada será

$$E(Y|X = \tilde{x}) = e^{\beta_0 + \boldsymbol{\beta}^\top \tilde{\mathbf{x}}} = e^{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}} \cdot e^{\beta_j}$$

- Fator de mudança de e^{β_j} no valor esperador
- Ficar atento à padronização dos preditores

Offset

- Modelagem de taxas: número de ocorrências num tempo $t > 0$

Offset

- Modelagem de taxas: número de ocorrências num tempo $t > 0$
- Se a coleta foi feita em tempos distintos t_i, então a contagem média será

$$E(Y_i|X = \mathbf{x}_i) = t_i \cdot \mu(\mathbf{x}_i),$$

em que $\mu(\mathbf{x}_i)$ é a taxa por unidade de tempo. Então,

$$\log E(Y_i|X = \mathbf{x}_i) = \log t_i + \log \mu(\mathbf{x}_i) = \log t_i + \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

Offset

- Modelagem de taxas: número de ocorrências num tempo $t > 0$
- Se a coleta foi feita em tempos distintos t_i, então a contagem média será

$$E(Y_i|X = \mathbf{x}_i) = t_i \cdot \mu(\mathbf{x}_i),$$

em que $\mu(\mathbf{x}_i)$ é a taxa por unidade de tempo. Então,

$$\log E(Y_i|X = \mathbf{x}_i) = \log t_i + \log \mu(\mathbf{x}_i) = \log t_i + \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$$

- $\log t_i$ é chamado **offset**, um termo conhecido que é usado no ajuste mas não é um preditor

Exemplo: simulação

- Simulação de um modelo log-linear com $n = 500$ e $p = 20$
- Preditores $\mathbf{x}_i \in \mathbb{R}^p$ gerados com variáveis aleatórias i.i.d. normal padrão e variável reposta gerada com $y_i|\mathbf{x}_i \sim Pois\left(e^{\beta_0+\boldsymbol{\beta}^\top \mathbf{x}_i}\right)$, em que $\beta_0 = 0$ e $\boldsymbol{\beta} = (-1, 0.5, -2, 1.5, 1, 0, \ldots, 0)^\top$
- Ajuste no `gmlnet` feito com a opção `family = 'poisson'`

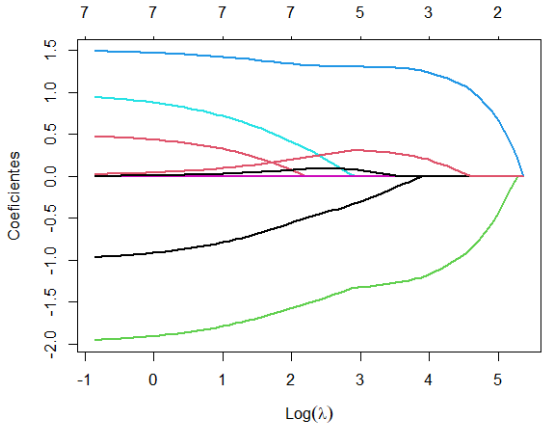

Figura: Trajetórias das estimativas para o modelo Poisson

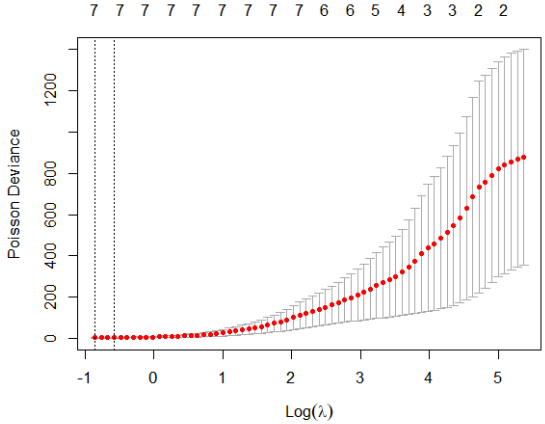

Figura: Deviance estimada via validação cruzada para diferentes λ's

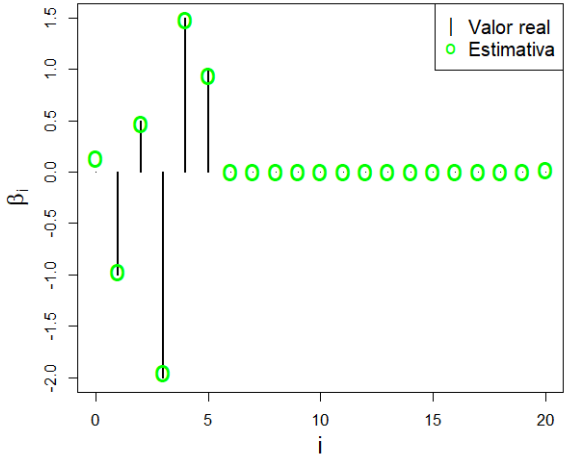

Figura: Valores verdadeiros dos parâmetros (linha preta) e estimativas (círculos verdes)

Exemplo: simulação com offset

- Simulação de um modelo log-linear com $n = 500$ e $p = 20$ com tempos de exposição distintos
- Preditores $\mathbf{x}_i \in \mathbb{R}^p$ gerados com v.a.'s i.i.d. normal padrão
- Tempos t_i i.i.d. com distribuição gama com parâmetros de forma 10 e escala 1
- Variável reposta gerada com $y_i|\mathbf{x}_i, t_i \sim Pois\left(t_i \cdot e^{\beta_0+\boldsymbol{\beta}^\top \mathbf{x}_i}\right)$, em que $\beta_0 = 0$ e $\boldsymbol{\beta} = (-1, 0.5, -2, 1.5, 1, 0, \ldots, 0)^\top$

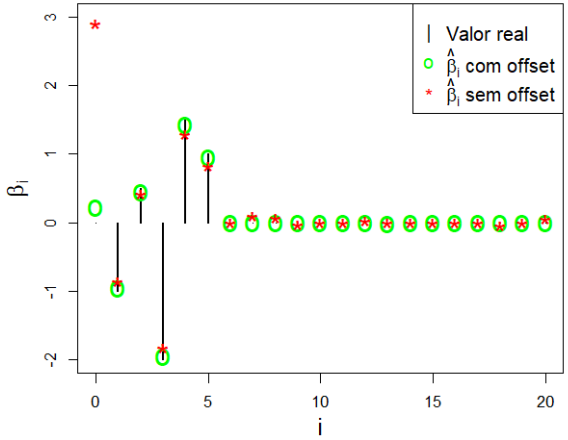

Figura: Valores verdadeiros dos parâmetros (linha preta), estimativas com offset (círculos verdes) e estimativas sem offset (asteriscos vermelhos)

Tabela: Estimativas com e sem utilizar o offset no ajuste

	Valor real	com offset	sem offset
β_0	0	0.2254	2.9037
β_1	-1	-0.9450	-0.8516
β_2	0.5	0.4547	0.4222
β_3	-2	-1.9488	-1.8202
β_4	1.5	1.4335	1.2920
β_5	1	0.9602	0.8321

Modelo de riscos proporcionais de Cox

- Análise do tempo até ocorrência de um evento de interesse
  - Análise de sobrevivência: tempo até morte/doença
  - Confiabilidade: tempo até falha/quebra

- Análise do tempo até ocorrência de um evento de interesse
  - Análise de sobrevivência: tempo até morte/doença
  - Confiabilidade: tempo até falha/quebra
- Presença de **dados censurados**, ou seja, tempo completo não é observado
- Censura à direita: tempo até evento é T, mas observamos $Y = \min\{T, C\}$, em que C é o tempo da censura

- Análise do tempo até ocorrência de um evento de interesse
  - Análise de sobrevivência: tempo até morte/doença
  - Confiabilidade: tempo até falha/quebra
- Presença de **dados censurados**, ou seja, tempo completo não é observado
- Censura à direita: tempo até evento é T, mas observamos $Y = \min\{T, C\}$, em que C é o tempo da censura
- Objetivo: estimar a **função de sobrevivência** $S(t|\mathbf{x}) = P(T > t|X = \mathbf{x})$

- $S(t)$ pose ser estimada de diferentes formas
  - Não paramétrica: Kaplan-Meier
  - Paramétrica: modelos de regressão exponencial, Weibull, etc.
  - Semi-paramétrica: **modelo de regressão de Cox**

- $S(t)$ pose ser estimada de diferentes formas
  - Não paramétrica: Kaplan-Meier
  - Paramétrica: modelos de regressão exponencial, Weibull, etc.
  - Semi-paramétrica: **modelo de regressão de Cox**

- Foco na função **taxa de falha**

$$h(t) = \lim_{\delta \to 0} \frac{P\big(Y \in (t, t+\delta) | Y \geq t\big)}{\delta} = \frac{f(t)}{S(t)},$$

em que $f(\cdot)$ denota a densidade de T

- Chance instantânea de falha no tempo t, dado que durou até o tempo t

Modelo de Cox

- Com a covariável $\mathbf{x} \in \mathbb{R}^p$, a função taxa de falha é tomada como

$$h(t) = h_0(t) \cdot e^{\boldsymbol{\beta}^\top \mathbf{x}},$$

em que $\boldsymbol{\beta} \in \mathbb{R}^p$ é um vetor de parâmetros e $h_0(t)$ é uma função não-negativa (função basal)

Modelo de Cox

- Com a covariável $\mathbf{x} \in \mathbb{R}^p$, a função taxa de falha é tomada como

$$h(t) = h_0(t) \cdot e^{\boldsymbol{\beta}^\top \mathbf{x}},$$

em que $\boldsymbol{\beta} \in \mathbb{R}^p$ é um vetor de parâmetros e $h_0(t)$ é uma função não-negativa (função basal)

- **Riscos proporcionais**: razão da taxa de falha de dois indivíduos i e j não depende do tempo t:

$$\frac{h_0(t) \cdot e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{h_0(t) \cdot e^{\boldsymbol{\beta}^\top \mathbf{x}_j}} = e^{\boldsymbol{\beta}^\top (\mathbf{x}_i - \mathbf{x}_j)}$$

Modelo de Cox

- Dados são $\{(\mathbf{x}_i, y_i, \delta_i)\}_{i=1}^n$, em que

$$\delta_i = \begin{cases} 1, & \text{se } y_i \text{ é tempo de falha} \\ 0, & \text{se } y_i \text{ é tempo de censura} \end{cases}$$

Modelo de Cox

- Dados são $\{(\mathbf{x}_i, y_i, \delta_i)\}_{i=1}^n$, em que

$$\delta_i = \begin{cases} 1, & \text{se } y_i \text{ é tempo de falha} \\ 0, & \text{se } y_i \text{ é tempo de censura} \end{cases}$$

- Sejam $y_1 < y_2 < \cdots < y_k$ os $k \leq n$ tempos de falhas distintos e $R(y_i)$ o conjunto de índices de observações sob risco (vivos e no experimento) até o tempo y_i

- Dado $R(y_i)$, a prob. da i-ésima observação falhar em y_i é

$$\frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{\sum_{j \in R(y_i)} e^{\boldsymbol{\beta}^\top \mathbf{x}_j}}$$

- Função de log-verossimilhança parcial com penalização ℓ_1:

$$Q(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \mathbb{I}\{\delta_i = 1\} \log \left(\frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{\sum_{j \in R(y_i)} e^{\boldsymbol{\beta}^\top \mathbf{x}_j}} \right) + \lambda \|\boldsymbol{\beta}\|_1$$

- Função de log-verossimilhança parcial com penalização ℓ_1:

$$Q(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \mathbb{I}\{\delta_i = 1\} \log \left(\frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{\sum_{j \in R(y_i)} e^{\boldsymbol{\beta}^\top \mathbf{x}_j}} \right) + \lambda \|\boldsymbol{\beta}\|_1$$

- Deviance é -2 vezes a função de log-verossimilhança parcial
- λ tipicamente é escolhido pela deviance mínima estimada via validação cruzada

Exemplo: dados de linfoma

- Tempo de vida de $n = 240$ pacientes dos quais 102 observações são censuradas
- Preditores são $p = 7399$ medidas de expressão genética
- Ajuste feito no `glmnet` com a opção `family='cox'`

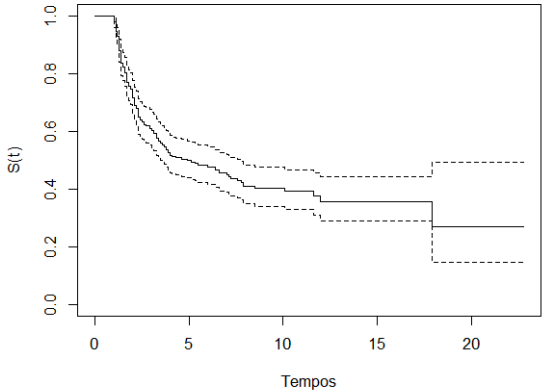

Figura: Estimativa da função da função de sobrevivência via Kaplan-Meier com intervalos de confiança de 95%.

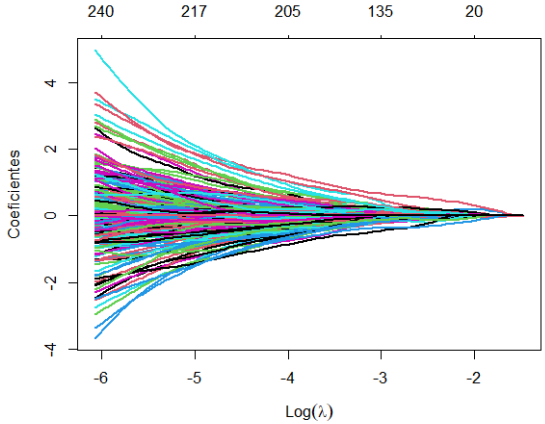

Figura: Trajetórias das estimativas para o modelo de Cox penalizado dos dados de linfoma.

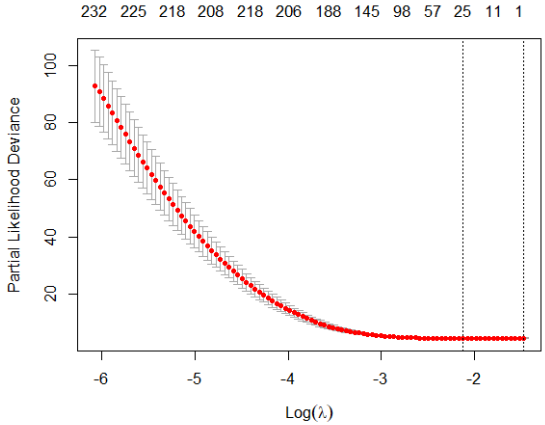

Figura: Deviance estimada via validação cruzada para diferentes λ's

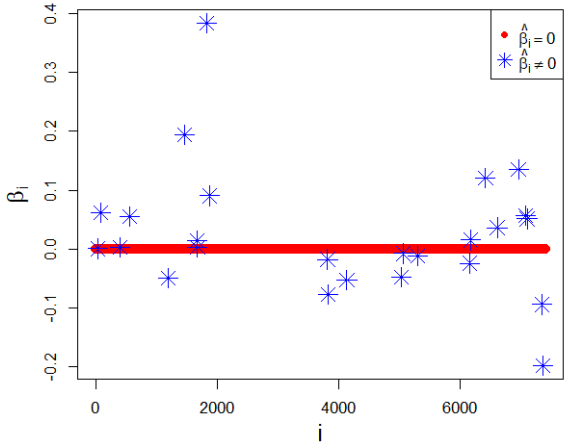

Figura: Estimativas do modelo de Cox penalizado, em vermelho para valores iguais a zero e em azul para valores diferentes de zero (25 no total).

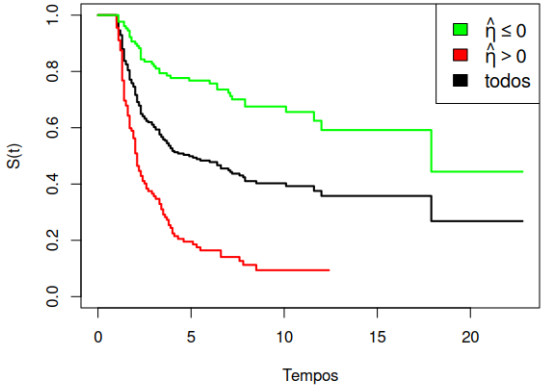

Figura: Kaplan-Meier para todas as observações (preto) e para grupos em que o preditor linear ajustado é positivo (verde) ou negativo (vermelho).

Referências

J. Fan, R. Li, C.-H. Zhang, and H. Zou. *Statistical Foundations of Data Science*. CRC Press, Boca Raton, 2020.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton, 2015.

A. G. Paula. Modelos de regressão: com apoio computacional. IME-USP, 2025.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Y. Zhang and D. N. Politis. Ridge regression revisited: Debiasing, thresholding and bootstrap. *The Annals of Statistics*, 50(3):1401–1422, 2022.