Inferência estatística em alta dimensão

Curso de Verão IMPA Tech – IDOR

Rodney Fonseca

Departamento de Estatística
Universidade Federal da Bahia

14 Jan 2026

Introdução

- Métodos com regularização ℓ_1: combinam seleção de variável e estimação de parâmetros
- Modelo selecionado com validação cruzada e validação realizada em um conjunto teste

- Métodos com regularização ℓ_1: combinam seleção de variável e estimação de parâmetros
- Modelo selecionado com validação cruzada e validação realizada em um conjunto teste
- **Inferência estatística**: como determinar a significância estatística das variáveis selecionadas?

- Métodos com regularização ℓ_1: combinam seleção de variável e estimação de parâmetros
- Modelo selecionado com validação cruzada e validação realizada em um conjunto teste
- **Inferência estatística**: como determinar a significância estatística das variáveis selecionadas?
- Métodos usuais (p-valores, intervalos de confiança/credibilidade) necessitam de adaptações em modelos regularizados

- Veremos alguns métodos de inferência em alta dimensão, com foco em **intervalos de confiança**
- Quantificar a incerteza de estimativas obtidas em modelos esparsos e sobre as variáveis selecionadas
- **Fonte:** Hastie et al. (2015, Cap. 6)

O lasso Bayesiano

- **Paradigma Bayesiano**: parâmetros são variáveis aleatórias com uma **distribuição a priori** que caracteriza o nosso conhecimento prévio sobre eles

- **Paradigma Bayesiano**: parâmetros são variáveis aleatórias com uma **distribuição a priori** que caracteriza o nosso conhecimento prévio sobre eles
- O método Bayesiano:
  1. Escolhemos a distribuição a prori $\pi(\theta)$ para θ

- **Paradigma Bayesiano**: parâmetros são variáveis aleatórias com uma **distribuição a priori** que caracteriza o nosso conhecimento prévio sobre eles

- O método Bayesiano:
  1. Escolhemos a distribuição a prori $\pi(\theta)$ para θ
  2. Escolhemos um modelo estatístico $f(x|\theta)$ que reflete o nosso conhecimento de X dado θ

- **Paradigma Bayesiano**: parâmetros são variáveis aleatórias com uma **distribuição a priori** que caracteriza o nosso conhecimento prévio sobre eles

- O método Bayesiano:
  1. Escolhemos a distribuição a prori $\pi(\theta)$ para θ
  2. Escolhemos um modelo estatístico $f(x|\theta)$ que reflete o nosso conhecimento de X dado θ
  3. Após observar $X_1, \ldots, X_n$, atualizamos nosso conhecimento sobre θ calculando a **distribuição a posteriori** $\pi(\theta|x_1, \ldots, x_n)$

- **Teorema de Bayes**: base do cálculo da posteriori

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta},$$

- **Teorema de Bayes**: base do cálculo da posteriori

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta},$$

- Sendo $L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ a função de verossimilhança para dados i.i.d. $\mathbf{x} = (x_1, \ldots, x_n)$, a posteriori será

$$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi(\theta)}{c_n} \propto L(\theta|\mathbf{x})\pi(\theta),$$

em que $c_n = \int L(\theta|\mathbf{x})\pi(\theta)d\theta$ é uma **constante de normalização**

- A distribuição a posteriori é usada para fazer inferência sobre θ
- Exemplos:
  - Estimação pontual através da moda a posteriori $\tilde{\theta} = \arg\max_\theta \pi(\theta|\mathbf{x})$

- A distribuição a posteriori é usada para fazer inferência sobre θ
- Exemplos:
  - Estimação pontual através da moda a posteriori $\tilde{\theta} = \arg\max_\theta \pi(\theta|\mathbf{x})$
  - Estimação pontual através da média a posteriori $\hat{\theta} = \int \theta \cdot \pi(\theta|\mathbf{x}) d\theta$
    $\hat{\theta}$ é o **estimador de Bayes** com função de perda quadrática

- A distribuição a posteriori é usada para fazer inferência sobre θ
- Exemplos:
  - Estimação pontual através da moda a posteriori $\tilde{\theta} = \arg\max_\theta \pi(\theta|\mathbf{x})$
  - Estimação pontual através da média a posteriori $\hat{\theta} = \int \theta \cdot \pi(\theta|\mathbf{x}) d\theta$
    $\hat{\theta}$ é o **estimador de Bayes** com função de perda quadrática
  - Estimação intervalar através de um intervalo C_α tal que

$$P(\theta \in C_\alpha|\mathbf{x}) = \int_{C_\alpha} \pi(\theta|\mathbf{x}) d\theta = 1 - \alpha$$

C_α é chamado **intervalo de credibilidade**

- No contexto de regressão normal linear, temos que $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$, em que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$
- Escolhemos distribuições a priori para os parâmetros β e σ^2

- No contexto de regressão normal linear, temos que $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$, em que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$
- Escolhemos distribuições a priori para os parâmetros β e σ^2
- Em alta dimensão, temos interesse em prioris de β que promovam uma **regularização das estimativas**

- No modelo linear normal ajustado via lasso, a função objetivo equivale a usar uma **priori Laplace** para os β's:

$$\mathbf{y}|\beta, \lambda, \sigma \sim N(\mathbf{X}\beta, \sigma^2 \boldsymbol{I}_n)$$

$$\beta|\lambda, \sigma \sim \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma}$$

- No modelo linear normal ajustado via lasso, a função objetivo equivale a usar uma **priori Laplace** para os β's:

$$\mathbf{y}|\beta, \lambda, \sigma \sim N(\mathbf{X}\beta, \sigma^2 \boldsymbol{I}_n)$$

$$\beta|\lambda, \sigma \sim \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma}$$

- Negativo da densidade a posteriori deste **modelo hierárquico**:

$$\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{\sigma}\|\beta\|_1 + c, \quad \text{com } c \text{ sendo uma constante}$$

- No modelo linear normal ajustado via lasso, a função objetivo equivale a usar uma **priori Laplace** para os β's:

$$\mathbf{y}|\beta, \lambda, \sigma \sim N(\mathbf{X}\beta, \sigma^2 \boldsymbol{I}_n)$$

$$\beta|\lambda, \sigma \sim \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma}$$

- Negativo da densidade a posteriori deste **modelo hierárquico**:

$$\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{\sigma}\|\beta\|_1 + c, \quad \text{com } c \text{ sendo uma constante}$$

- O estimador lasso equivale à **moda a posteriori**

- Com toda uma distribuição a posteriori $\pi(\theta|\mathbf{x})$ temos:
  - Estimativas pontuais
  - Estimativas intervalares
  - Erros padrão
  - Quantis

- Com toda uma distribuição a posteriori $\pi(\theta|\mathbf{x})$ temos:
  - Estimativas pontuais
  - Estimativas intervalares
  - Erros padrão
  - Quantis

- Mais flexibilidade para lidar com hiperparâmetros (σ, λ, etc) incluindo-os no modelo com uma distribuição a priori

- Com toda uma distribuição a posteriori $\pi(\theta|\mathbf{x})$ temos:
  - Estimativas pontuais
  - Estimativas intervalares
  - Erros padrão
  - Quantis

- Mais flexibilidade para lidar com hiperparâmetros (σ, λ, etc) incluindo-os no modelo com uma distribuição a priori

- **Desvantagem:** a distribuição a posteriori tipicamente precisa ser calculada numericamente (amostrador de Gibbs, MCMC, etc.), o que pode ter uma custo computacional alto

Exemplo: dados de diabetes

- Dados de $n = 442$ pacientes com uma medida de progresso da diabetes
- As variáveis explicativas são:
  - X_1: sexo
  - X_2: idade do paciente
  - X_3: índice de massa corporal
  - X_4: pressão sanguínea média
  - $X_5, \ldots, X_{10}$: medidas sanguíneas
- Todas medidas foram normalizadas para a análise

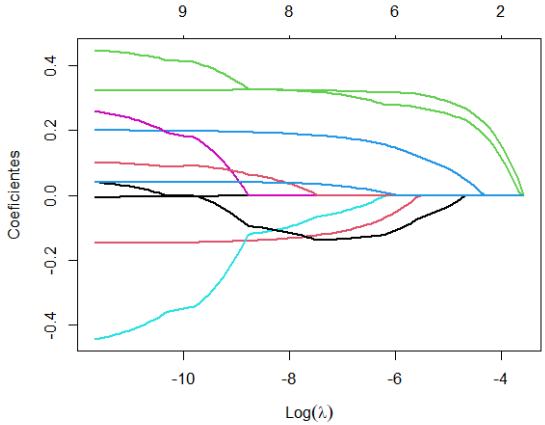

Figura: Trajetória das estimativas do lasso para diferentes penalizações λ

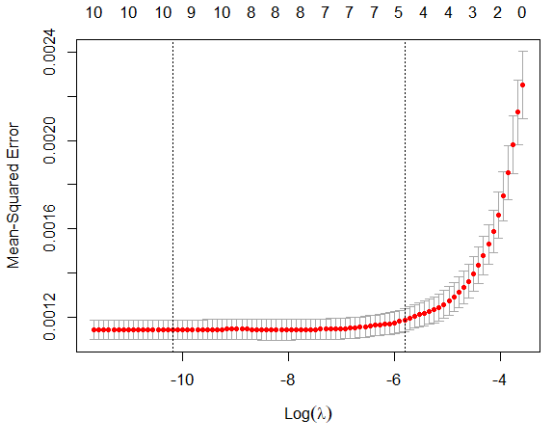

Figura: EQM estimado via validação cruzada para diferentes λ's

Tabela: Estimativas de $\beta_1, \ldots, \beta_{10}$ pelo lasso com $\lambda = 0.0029$

	Lasso
β_1	0
β_2	-0.0317
β_3	0.3147
β_4	0.1362
β_5	0
β_6	0
β_7	-0.0940
β_8	0
β_9	0.2762
β_{10}	0

Lasso Bayesiano no R

- Ajuste feito através da função `blasso` do pacote de mesmo nome
- Distribuições a posteriori são obtidas numericamente gerando valores aleatórios delas a partir de MCMC
- Precisamos escolher o tamanho da cadeia de MCMC e número de lags que separam as observações mantidas

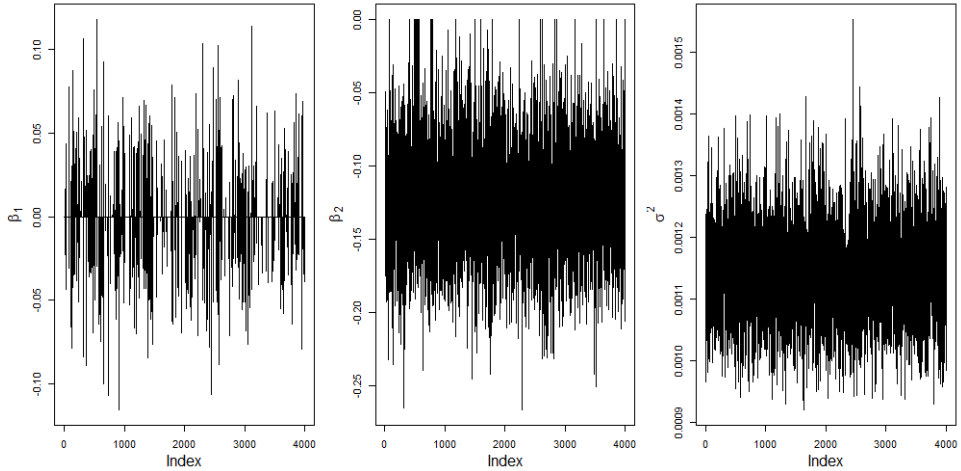

Figura: Cadeias de MCMC de 4000 valores gerados para β_1, β_2 e σ^2

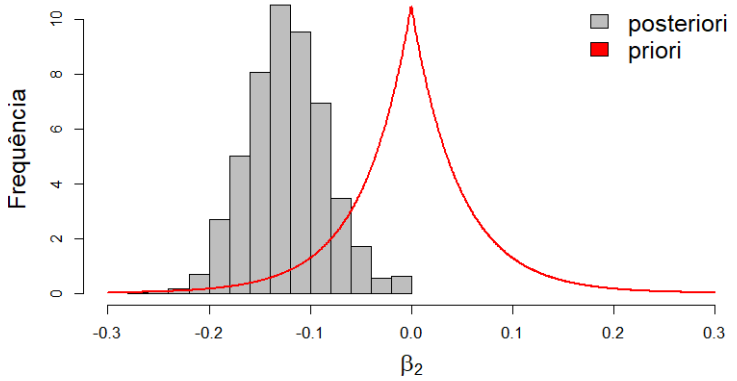

Figura: Distribuição a posteriori e ilustração da distribuição a priori Laplace com média zero e parâmetro de escala $DP(\mathbf{y})$, o desvio padrão da resposta $\mathbf{y}$

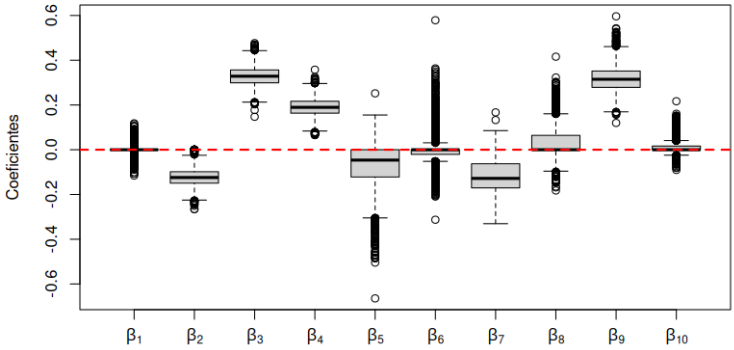

Figura: Boxplot dos valores dos parâmetros gerados via MCMC

Tabela: Estimativas de $\beta_1, \ldots, \beta_{10}$ usando o Lasso usual (frequentista) e estimativas a posteriori: média, desvio padrão (DP) e intervalo de 95% credibilidade (IC)

	Lasso	Média post.	DP post.	IC 95% post.
β_1	0	-0.0002	0.0158	(-0.0408, 0.0391)
β_2	-0.0317	-0.1231	0.0396	(-0.1952, -0.0415)
β_3	0.3147	0.3278	0.0418	(0.2461 , 0.4088)
β_4	0.1362	0.1901	0.0397	(0.1132 , 0.2685)
β_5	0	-0.0706	0.0880	(-0.2889 , 0.0155)
β_6	0	-0.0069	0.0589	(-0.1223 , 0.1485)
β_7	-0.0940	-0.1160	0.0744	(-0.2395 , 0.0000)
β_8	0	0.0350	0.0691	(-0.0593 , 0.2070)
β_9	0.2762	0.3168	0.0557	(0.2134 , 0.4384)
β_{10}	0	0.0133	0.0289	(-0.0145, 0.0924)

O bootstrap

- Seja $T_n = g(X_1, \ldots, X_n)$ uma estatística de dados i.i.d. com função de distribuição F
- Obter certas propriedades (viés, variância, etc) de T_n requer conhecer a sua distribuição, algo nem sempre possível

- Seja $T_n = g(X_1, \ldots, X_n)$ uma estatística de dados i.i.d. com função de distribuição F
- Obter certas propriedades (viés, variância, etc) de T_n requer conhecer a sua distribuição, algo nem sempre possível
- **Alternativas:**
  1. distribuição assintótica de T_n

- Seja $T_n = g(X_1, \ldots, X_n)$ uma estatística de dados i.i.d. com função de distribuição F
- Obter certas propriedades (viés, variância, etc) de T_n requer conhecer a sua distribuição, algo nem sempre possível
- **Alternativas:**
  1. distribuição assintótica de T_n
  2. métodos numéricos para aproximar a distribuição de T_n

- Seja $T_n = g(X_1, \ldots, X_n)$ uma estatística de dados i.i.d. com função de distribuição F
- Obter certas propriedades (viés, variância, etc) de T_n requer conhecer a sua distribuição, algo nem sempre possível
- **Alternativas:**
  1. distribuição assintótica de T_n
  2. métodos numéricos para aproximar a distribuição de T_n
- **Bootstrap:** Método numérico muito usado para estimar erro padrão e intervalo de confiança

Exemplo: variância de T_n

- Note que $Var(T_n)$ depende de F
  - Se $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ e $T_n = \bar{X}_n$, então $Var(T_n) = \sigma^2/n$, em que $\sigma^2 = \int (x-\mu)^2 dF(x)$ e $\mu = \int x dF(x)$

Exemplo: variância de T_n

- Note que $Var(T_n)$ depende de F
  - Se $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ e $T_n = \bar{X}_n$, então $Var(T_n) = \sigma^2/n$, em que $\sigma^2 = \int (x-\mu)^2 dF(x)$ e $\mu = \int x dF(x)$

- Ideia do **bootstrap não-paramétrico**:
  1. Aproximar $Var_F(T_n)$ por $Var_{\hat{F}_n}(T_n)$, em que $\hat{F}_n$ é a função de distribuição empírica dos dados
  2. Estimar $Var_{\hat{F}_n}(T_n)$ gerando amostras aleatórias a partir de $\hat{F}_n$

Exemplo: variância de T_n

- Note que $Var(T_n)$ depende de F
  - Se $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ e $T_n = \bar{X}_n$, então $Var(T_n) = \sigma^2/n$, em que $\sigma^2 = \int (x-\mu)^2 dF(x)$ e $\mu = \int x dF(x)$
- Ideia do **bootstrap não-paramétrico**:
  1. Aproximar $Var_F(T_n)$ por $Var_{\hat{F}_n}(T_n)$, em que $\hat{F}_n$ é a função de distribuição empírica dos dados
  2. Estimar $Var_{\hat{F}_n}(T_n)$ gerando amostras aleatórias a partir de $\hat{F}_n$
- **Vantagem:** estimamos $Var_F(T_n)$ sem precisar deduzir a distribuição de T_n

- A **função de distribuição empírica** é definida como

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\},$$

em que $\mathbb{I}(\cdot)$ é uma função indicadora

- A **função de distribuição empírica** é definida como

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\},$$

em que $\mathbb{I}(\cdot)$ é uma função indicadora

- $\hat{F}_n$ é uma distribuição uniforme discreta que atribui probabilidade $1/n$ para cada X_i dos dados $X_1, \ldots, X_n$

- A **função de distribuição empírica** é definida como

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\},$$

em que $\mathbb{I}(\cdot)$ é uma função indicadora

- $\hat{F}_n$ é uma distribuição uniforme discreta que atribui probabilidade $1/n$ para cada X_i dos dados $X_1, \ldots, X_n$
- Gerar amostras de $\hat{F}_n$ equivale a sortear valores de $\{X_1, \ldots, X_n\}$ **com reposição**

- Ilustração do bootstrap não-paramétrico:

$$\text{Mundo real } F \Rightarrow X_1, \ldots, X_n \Rightarrow T_n = g(X_1, \ldots, X_n)$$

- Ilustração do bootstrap não-paramétrico:

$$\text{Mundo real } F \Rightarrow X_1, \ldots, X_n \Rightarrow T_n = g(X_1, \ldots, X_n)$$

$$\text{Mundo bootstrap } \hat{F}_n \Rightarrow X_1^*, \ldots, X_n^* \Rightarrow T_n^* = g(X_1^*, \ldots, X_n^*)$$

- Ilustração do bootstrap não-paramétrico:

$$\text{Mundo real } F \Rightarrow X_1, \ldots, X_n \Rightarrow T_n = g(X_1, \ldots, X_n)$$

$$\text{Mundo bootstrap } \hat{F}_n \Rightarrow X_1^*, \ldots, X_n^* \Rightarrow T_n^* = g(X_1^*, \ldots, X_n^*)$$

- A amostra $X_1^*, \ldots, X_n^*$ de $\hat{F}_n$ é obtida sorteando **com reposição** n valores de $\{X_1, \ldots, X_n\}$
- Gerando vários T_n^*'s independentes, **podemos estimar a distribuição** de T_n

Erro padrão boostrap

- Estimar $Var(T_n)$ via bootstrap:
  1. Tome uma amostra $X_1^*, \ldots, X_n^*$ de $\{X_1, \ldots, X_n\}$ com reposição;
  2. Calcule $T_n^* = g(X_1^*, \ldots, X_n^*)$
  3. Repita os passos 1 e 2 por $B - 1$ vezes, gerando $T_{n1}^*, \ldots, T_{nB}^*$
  4. Calcule $V_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left(T_{nb}^* - \frac{1}{B} \sum_{k=1}^{B} T_{nk}^* \right)^2$
- O **erro padrão bootstrap** será $\sqrt{V_{\text{boot}}}$

Intervalo de confiança bootstrap

- Construir um intervalo de $(1-\alpha)100\%$ de confiança para T_n:
  1. Tome uma amostra $X_1^*, \ldots, X_n^*$ de $\{X_1, \ldots, X_n\}$ com reposição;
  2. Calcule $T_n^* = g(X_1^*, \ldots, X_n^*)$
  3. Repita os passos 1 e 2 por $B-1$ vezes, gerando $T_{n1}^*, \ldots, T_{nB}^*$
  4. Calcule $C_n = [T_{\alpha/2}^*, T_{1-\alpha/2}^*]$ em que T_γ^* é o quantil $\gamma 100\%$ de $T_{n1}^*, \ldots, T_{nB}^*$
- C_n é chamado **IC bootstrap percentil**

- O mesmo princípio pode ser aplicado para estimar outras quantidades de T_n

- O mesmo princípio pode ser aplicado para estimar outras quantidades de T_n
- A **distribuição boostrap** é usada de forma parecida à distribuição a posteriori na abordagem Bayesiana

- O mesmo princípio pode ser aplicado para estimar outras quantidades de T_n
- A **distribuição boostrap** é usada de forma parecida à distribuição a posteriori na abordagem Bayesiana
- **Observação:** no caso de regressão, a reamostragem é feita com os pares de resposta & preditor $(y_1, \mathbf{X}_1), \ldots, (y_n, \mathbf{X}_n)$, em que $\mathbf{X}_i \in \mathbb{R}^p$ é o vetor de preditores

Estimadores regularizados/penalizados

- Geralmente não conhecemos a distribuição (nem assintótica) de estimadores pontuais $\hat{\beta}$ de modelos regularizados como o lasso

Estimadores regularizados/penalizados

- Geralmente não conhecemos a distribuição (nem assintótica) de estimadores pontuais $\hat{\beta}$ de modelos regularizados como o lasso
- Método bootstrap é útil para estimar a distribuição de $\hat{\beta}$ e quantificar a incerteza sobre estimativas regularizadas

Estimadores regularizados/penalizados

- Geralmente não conhecemos a distribuição (nem assintótica) de estimadores pontuais $\hat{\beta}$ de modelos regularizados como o lasso
- Método bootstrap é útil para estimar a distribuição de $\hat{\beta}$ e quantificar a incerteza sobre estimativas regularizadas
- Boostrap também auxilia a estimar a distribuição da regularização λ escolhida

Lasso bootstrap

- Seja $\hat{\beta}(\lambda)$ uma estimativa lasso obtida com $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$
  1. Tome uma amostra $(y_1^*, \mathbf{X}_1^*), \ldots, (y_n^*, \mathbf{X}_n^*)$ sorteando pares de $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$ com reposição;
  2. Escolha λ^* (via CV, por exemplo) e calcule $\hat{\beta}(\lambda^*)$ usando $\{(y_i^*, \mathbf{X}_i^*)\}_{i=1}^n$
  3. Repita os passos 1 e 2 por $B-1$ vezes, gerando $\lambda_1^*, \ldots, \lambda_B^*$ e $\hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_B^*$
  4. Calcule as estimativas bootstrap de interesse

Lasso bootstrap

- Seja $\hat{\beta}(\lambda)$ uma estimativa lasso obtida com $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$
  1. Tome uma amostra $(y_1^*, \mathbf{X}_1^*), \ldots, (y_n^*, \mathbf{X}_n^*)$ sorteando pares de $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$ com reposição;
  2. Escolha λ^* (via CV, por exemplo) e calcule $\hat{\beta}(\lambda^*)$ usando $\{(y_i^*, \mathbf{X}_i^*)\}_{i=1}^n$
  3. Repita os passos 1 e 2 por $B-1$ vezes, gerando $\lambda_1^*, \ldots, \lambda_B^*$ e $\hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_B^*$
  4. Calcule as estimativas bootstrap de interesse
- **Atenção:** o mesmo procedimento deve ser realizado em cada b-ésima repetição

Exemplo: dados de diabetes

- Dados de $n = 442$ pacientes com uma medida de progresso da diabetes, contendo $p = 10$ variáveis explicativas
- As variáveis explicativas são:
  - X_1: sexo
  - X_2: idade do paciente
  - X_3: índice de massa corporal
  - X_4: pressão sanguínea média
  - $X_5, \ldots, X_{10}$: medidas sanguíneas
- Análise anterior com o lasso selecionou X_2, X_3, X_4, X_7 e X_9
- Aplicamos bootstrap com $B = 300$ réplicas

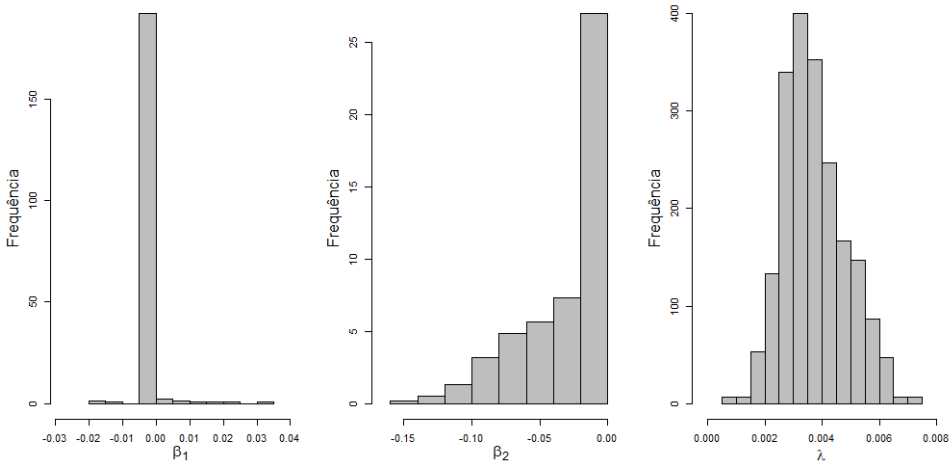

Figura: Histograma das estimativas bootstrap de β_1^*, β_2^* e λ^*

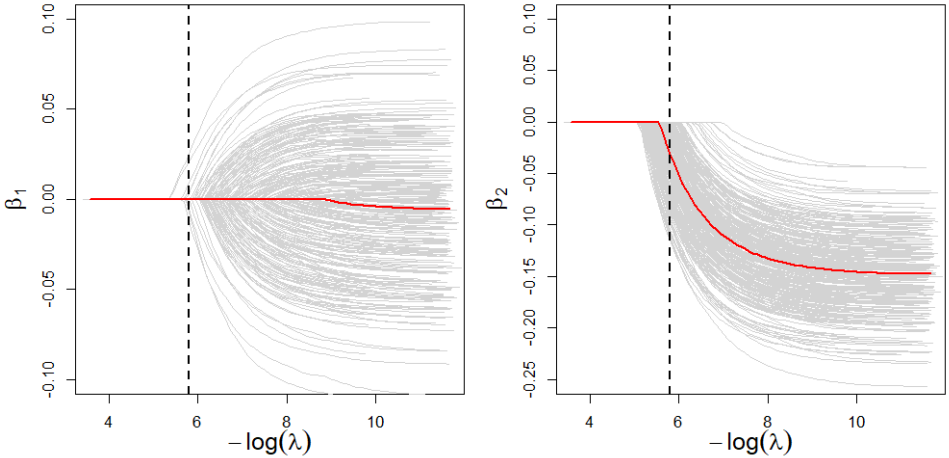

Figura: Trajetórias das estimativas bootstrap de β_1^* e β_2^* vs. λ^*

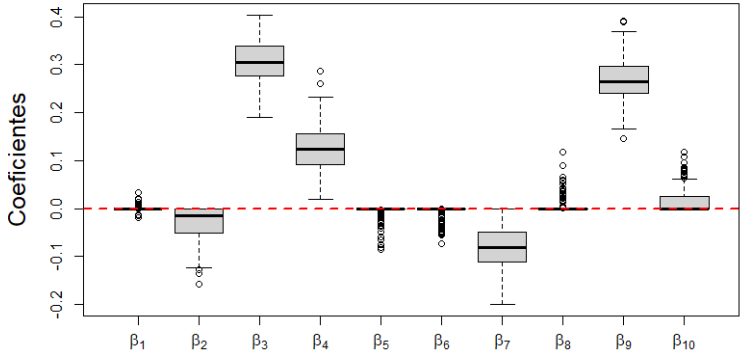

Figura: Boxplots das estimativas bootstrap

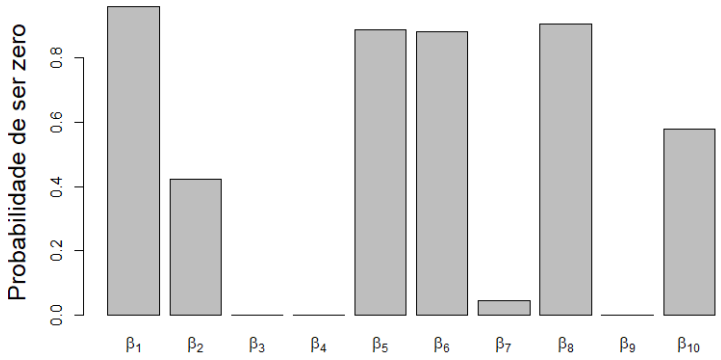

Figura: Proporção de estimativas bootstrap que foram zero para cada parâmetro

Tabela: Estimativas de $\beta_1, \ldots, \beta_{10}$ usando o Lasso com λ_{CV}, erro padrão bootstrap (EP) e intervalo de 95% de confiança bootstrap (IC)

	Lasso	EP	IC 95%
β_1	0	0.00	[0.00, 0.00]
β_2	-0.0317	0.03	[-0.10, 0.00]
β_3	0.3147	0.04	[0.22, 0.39]
β_4	0.1362	0.04	[0.04, 0.20]
β_5	0	0.01	[-0.04, 0.00]
β_6	0	0.01	[-0.04, 0.00]
β_7	-0.0940	0.04	[-0.17, -0.00]
β_8	0	0.01	[0.00, 0.03]
β_9	0.2762	0.05	[0.19, 0.35]
β_{10}	0	0.02	[0.00, 0.07]

Lasso deviesado

- **Distribuição assintótica** do estimador de mínimos quadrados no modelo linear:

$$\hat{\beta}_{mqo} \sim N\big(\beta^*, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}\big), \quad n > p, \ n \text{ grande},$$

em que β^* é o vetor de coeficientes de regressão verdadeiros

- **Distribuição assintótica** do estimador de mínimos quadrados no modelo linear:

$$\hat{\beta}_{mqo} \sim N\big(\beta^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big), \quad n > p,\ n \text{ grande},$$

em que β^* é o vetor de coeficientes de regressão verdadeiros

- **Inferência aproximada** usando a distribuição assintótica
  - IC para β_j^*: $\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \sqrt{v_j}\hat{\sigma}$, em que v_j é o j-ésimo elemento da diagonal de $(\mathbf{X}^\top\mathbf{X})^{-1}$

- **Distribuição assintótica** do estimador de mínimos quadrados no modelo linear:

$$\hat{\beta}_{mqo} \sim N\big(\beta^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big), \quad n > p,\ n \text{ grande},$$

em que β^* é o vetor de coeficientes de regressão verdadeiros

- **Inferência aproximada** usando a distribuição assintótica
  - IC para β_j^*: $\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \sqrt{v_j}\hat{\sigma}$, em que v_j é o j-ésimo elemento da diagonal de $(\mathbf{X}^\top\mathbf{X})^{-1}$
  - Teste t para $H_0 : \beta_j^* = 0$ vs. $H_1 : \beta_j^* \neq 0$: $t_j = \frac{\hat{\beta}_j}{\sqrt{v_j}\hat{\sigma}}$, rejeitando H_0 ao nível α se $|t_j| > t_{n-p,1-\alpha/2}$

- Problema com **estimador lasso** $\hat{\beta}_\lambda$: viesado e não conhecemos a sua distribuição para grandes amostras

- Problema com **estimador lasso** $\hat{\beta}_\lambda$: viesado e não conhecemos a sua distribuição para grandes amostras
- O **lasso deviesado** consiste em aplicar uma correção de viés a $\hat{\beta}_\lambda$ que permita deduzir a sua distribuição assintótica

- Problema com **estimador lasso** $\hat{\beta}_\lambda$: viesado e não conhecemos a sua distribuição para grandes amostras
- O **lasso deviesado** consiste em aplicar uma correção de viés a $\hat{\beta}_\lambda$ que permita deduzir a sua distribuição assintótica
- Tal distribuição pode ser utilizada para fazer inferência aproximada para β^*

- Considere o modelo linear normal

$$\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_p)$$

- Se $n > p$, o estimador de MQO pode ser escrito como

$$\hat{\beta}_{mqo} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \beta^* + \frac{1}{\sqrt{n}} \left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}^\top}{\sqrt{n}} \boldsymbol{\epsilon}$$

- Considere o modelo linear normal

$$\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_p)$$

- Se $n > p$, o estimador de MQO pode ser escrito como

$$\hat{\beta}_{mqo} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \beta^* + \frac{1}{\sqrt{n}} \underbrace{\left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}^\top}{\sqrt{n}} \boldsymbol{\epsilon}}_{N(\mathbf{0}, \sigma^2 (\mathbf{X}^\top \mathbf{X}/n)^{-1}), \text{ dado } \mathbf{X}}$$

- Considere o modelo linear normal

$$\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_p)$$

- Se $n > p$, o estimador de MQO pode ser escrito como

$$\hat{\beta}_{mqo} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \beta^* + \frac{1}{\sqrt{n}} \underbrace{\left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}^\top}{\sqrt{n}} \boldsymbol{\epsilon}}_{N(\mathbf{0}, \sigma^2 (\mathbf{X}^\top \mathbf{X}/n)^{-1}), \text{ dado } \mathbf{X}}$$

- Logo, $\hat{\beta}_{mqo} | \mathbf{X} \sim N\left(\beta^*, \sigma^2 (\mathbf{X}^\top \mathbf{X}/n)^{-1}\right)$

- O **lasso deviesado** é dado por

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{n}\hat{\Theta}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda),$$

em que $\hat{\Theta}$ é uma aproximação da inversa de $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$

- O **lasso deviesado** é dado por

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{n}\hat{\Theta}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda),$$

em que $\hat{\Theta}$ é uma aproximação da inversa de $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$

- Podemos reescrever $\hat{\beta}^d$ como

$$\hat{\beta}^d = \beta^* + \frac{1}{\sqrt{n}}\hat{\Theta}\frac{\mathbf{X}^\top}{\sqrt{n}}\boldsymbol{\epsilon} + (\boldsymbol{I}_p - \hat{\Theta}\hat{\Sigma})(\hat{\beta}_\lambda - \beta^*)$$

- O **lasso deviesado** é dado por

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{n}\hat{\Theta}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda)$$

em que $\hat{\Theta}$ é uma aproximação da inversa de $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$

- Podemos reescrever $\hat{\beta}^d$ como

$$\hat{\beta}^d = \beta^* + \frac{1}{\sqrt{n}}\underbrace{\hat{\Theta}\frac{\mathbf{X}^\top}{\sqrt{n}}\boldsymbol{\epsilon}}_{N(\mathbf{0},\sigma^2\hat{\Theta}\hat{\Sigma}\hat{\Theta}^\top),\text{ dado }\mathbf{X}} + \underbrace{(\boldsymbol{I}_p - \hat{\Theta}\hat{\Sigma})(\hat{\beta}_\lambda - \beta^*)}_{\hat{\Delta}}$$

- O **lasso deviesado** é dado por

$$\hat{\beta}^d = \hat{\beta}_\lambda + \frac{1}{n}\hat{\Theta}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda)$$

em que $\hat{\Theta}$ é uma aproximação da inversa de $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$

- Podemos reescrever $\hat{\beta}^d$ como

$$\hat{\beta}^d = \beta^* + \frac{1}{\sqrt{n}}\underbrace{\hat{\Theta}\frac{\mathbf{X}^\top}{\sqrt{n}}\boldsymbol{\epsilon}}_{N(\mathbf{0},\sigma^2\hat{\Theta}\hat{\Sigma}\hat{\Theta}^\top),\text{ dado }\mathbf{X}} + \underbrace{(\boldsymbol{I}_p - \hat{\Theta}\hat{\Sigma})(\hat{\beta}_\lambda - \beta^*)}_{\hat{\Delta}}$$

- Se $\hat{\Delta} \approx \mathbf{0}$, então $\hat{\beta}^d$ é aproximadamente $N(\beta^*, \sigma^2\hat{\Theta}\hat{\Sigma}\hat{\Theta}^\top)$

- Se $n > p$, tomando $\hat{\Theta} = \hat{\Sigma}^{-1}$ obtemos $\hat{\beta}^d = \hat{\beta}_{mqo}$
- **Desafio** quando $p > n$: obter $\hat{\Theta}$ tal que o viés remanescente de $\hat{\beta}^d$ seja pequeno, ou seja,

$$\|\hat{\Delta}\| = \|(\boldsymbol{I}_p - \hat{\Theta}\hat{\Sigma})(\hat{\beta}_\lambda - \beta^*)\| \approx 0$$

- Se $n > p$, tomando $\hat{\Theta} = \hat{\Sigma}^{-1}$ obtemos $\hat{\beta}^d = \hat{\beta}_{mqo}$
- **Desafio** quando $p > n$: obter $\hat{\Theta}$ tal que o viés remanescente de $\hat{\beta}^d$ seja pequeno, ou seja,

$$\|\hat{\Delta}\| = \|(\boldsymbol{I}_p - \hat{\Theta}\hat{\Sigma})(\hat{\beta}_\lambda - \beta^*)\| \approx 0$$

- Possíveis abordagens:
  - Calcular $\hat{\Theta}$ tal que $\|\hat{\Sigma}\hat{\Theta} - \boldsymbol{I}_p\|$ seja pequeno (Javanmard and Montanari, 2014b)

- Se $n > p$, tomando $\hat{\Theta} = \hat{\Sigma}^{-1}$ obtemos $\hat{\beta}^d = \hat{\beta}_{mqo}$
- **Desafio** quando $p > n$: obter $\hat{\Theta}$ tal que o viés remanescente de $\hat{\beta}^d$ seja pequeno, ou seja,

$$\|\hat{\Delta}\| = \|(\boldsymbol{I}_p - \hat{\Theta}\hat{\Sigma})(\hat{\beta}_\lambda - \beta^*)\| \approx 0$$

- Possíveis abordagens:
  - Calcular $\hat{\Theta}$ tal que $\|\hat{\Sigma}\hat{\Theta} - \boldsymbol{I}_p\|$ seja pequeno (Javanmard and Montanari, 2014b)
  - Calcular $\hat{\Theta}$ como uma **estimativa esparsa da matriz de precisão** Σ^{-1} (van de Geer et al., 2014)

Exemplo: dados de diabetes

- Dados de $n = 442$ pacientes com uma medida de progresso da diabetes, contendo 10 variáveis explicativas
- Ajustaremos um modelo incluindo termos quadráticos e interações, com total de $p = 64$ covariáveis
- Lista das covariáveis: sexo (X_1), idade (X_2), IMC (X_3), pressão (X_4), medidas sanguíneas ($X_5, \ldots, X_{10}$), termos quadráticos e interações

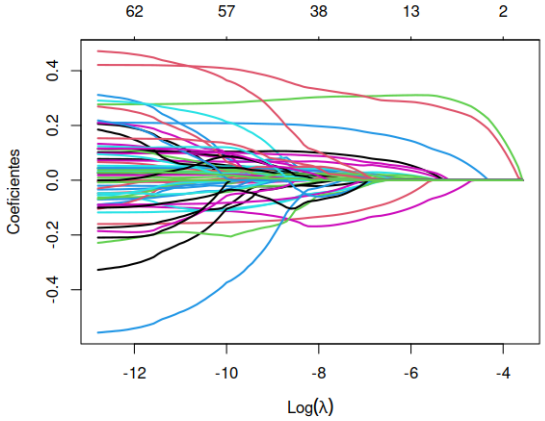

Figura: Trajetória das estimativas lasso do modelo quadrático para diferentes penalizações λ

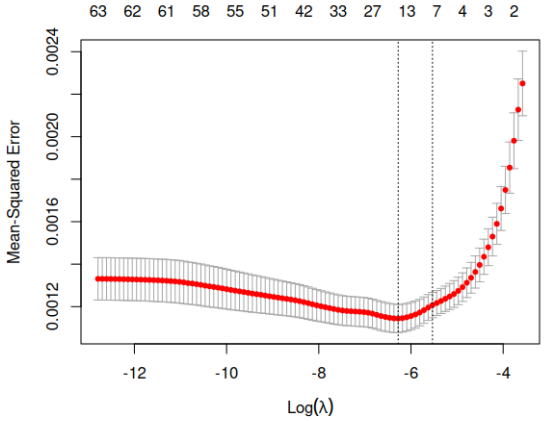

Figura: EQM estimado via validação cruzada do modelo quadrático para diferentes λ's

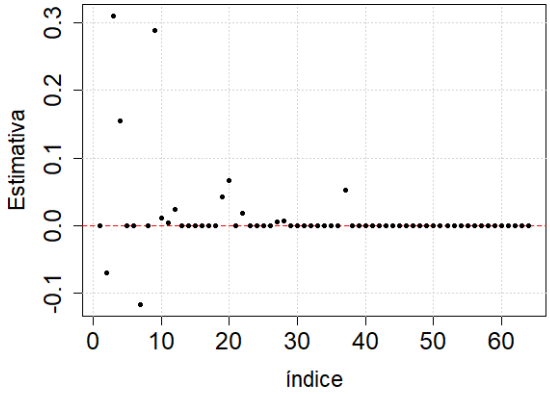

Figura: Estimativas do modelo quadrático utilizando lasso com penalização $\lambda_{CV} = 0.0018$

Tabela: Dez maiores estimativas (em valor absoluto) do modelo quadrático usando lasso e $\lambda_{CV} = 0.0018$

Parâmetro	Lasso
β_4	0.31
β_{10}	0.29
β_5	0.16
β_8	-0.12
β_3	-0.07
β_{21}	0.07
β_{38}	0.05
β_{20}	0.04
β_{13}	0.02
β_{23}	0.02

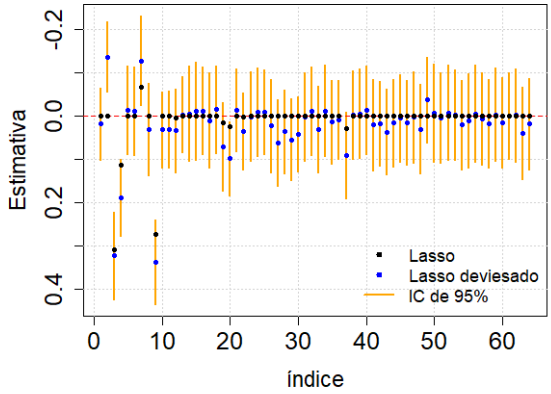

Figura: Estimativas do modelo quadrático utilizando lasso deviesado com IC's de 95% de confiança

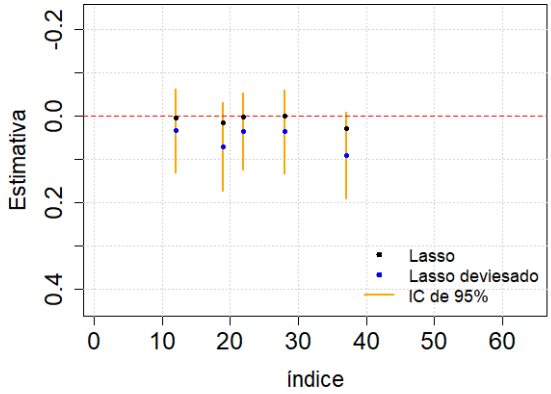

Figura: Estimativas lasso que são diferentes de zero mas cujo IC de 95% obtido com lasso deviesado contém zero

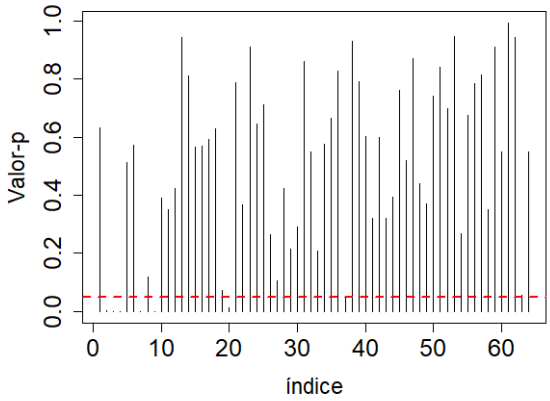

Figura: Valores-p paras as estimativas obtidas via lasso deviesado. Linha vermelha marca o nível $\alpha = 0.05$

Tabela: Dez maiores estimativas do modelo quadrático usando lasso e $\lambda_{CV} = 0.0018$, estimativas correspondentes pelo lasso deviesado e IC's de 95%

Parâmetro	Lasso	Dev. lasso	IC
β_4	0.31	0.19	(0.10, 0.28)
β_{10}	0.29	0.03	(-0.06, 0.12)
β_5	0.16	-0.01	(-0.11, 0.09)
β_8	-0.12	0.03	(-0.07, 0.14)
β_3	-0.07	0.32	(0.22, 0.42)
β_{21}	0.07	-0.01	(-0.11, 0.08)
β_{38}	0.05	-0.00	(-0.10, 0.10)
β_{20}	0.04	0.10	(0.01, 0.18)
β_{13}	0.02	-0.00	(-0.09, 0.08)
β_{23}	0.02	0.00	(-0.10, 0.11)

Observações

- Conseguimos fazer inferência com o lasso deviesado, mas as estimativas pontuais não são esparsas

Observações

- Conseguimos fazer inferência com o lasso deviesado, mas as estimativas pontuais não são esparsas
- Também existe estimadores deviesados para lasso MLG, caso em que a correção de viés envolve outra função objetivo $\mathcal{L}(\beta)$ e sua matriz Hessiana

Observações

- Conseguimos fazer inferência com o lasso deviesado, mas as estimativas pontuais não são esparsas
- Também existe estimadores deviesados para lasso MLG, caso em que a correção de viés envolve outra função objetivo $\mathcal{L}(\beta)$ e sua matriz Hessiana
- Testes de hipóteses múltiplos necessitam ajustes no nível de significância (e.g., Bonferroni) ou correções que controlem a taxa de rejeições falsas (e.g., FDR)

Referências

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

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1): 2869–2909, 2014b.

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.