Aprendizagem estatística em altas dimensões

Florencia Leonardi

Conteúdo

- Seleção de variáveis
- Aproximação do erro fora da amostra - BIC
- Melhor subconjunto, seleção progressiva e seleção regressiva
- Regularização, RIDGE, LASSO e Elastic Net

Ajuste versus complexidade

$f\colon [-1,1] \to \mathbb{R}$

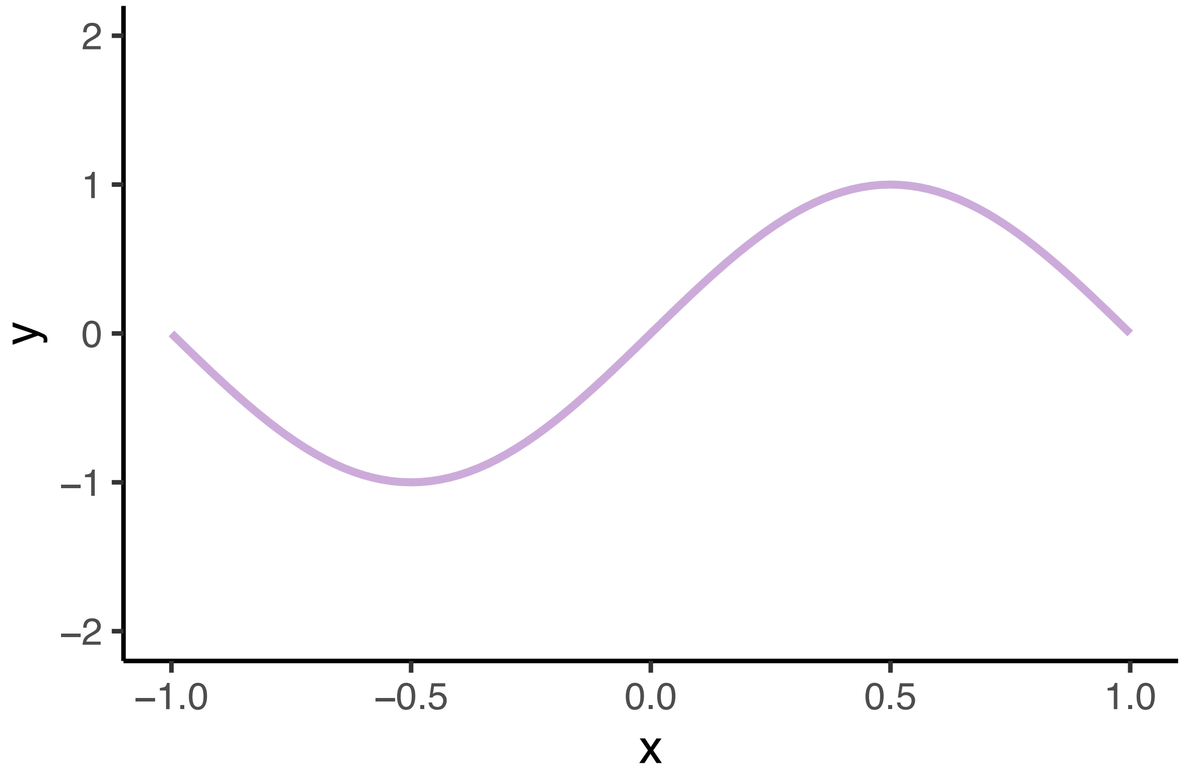
$f(x) = \sin(\pi x)$

$\epsilon = 0 \qquad\qquad y = f(x)$

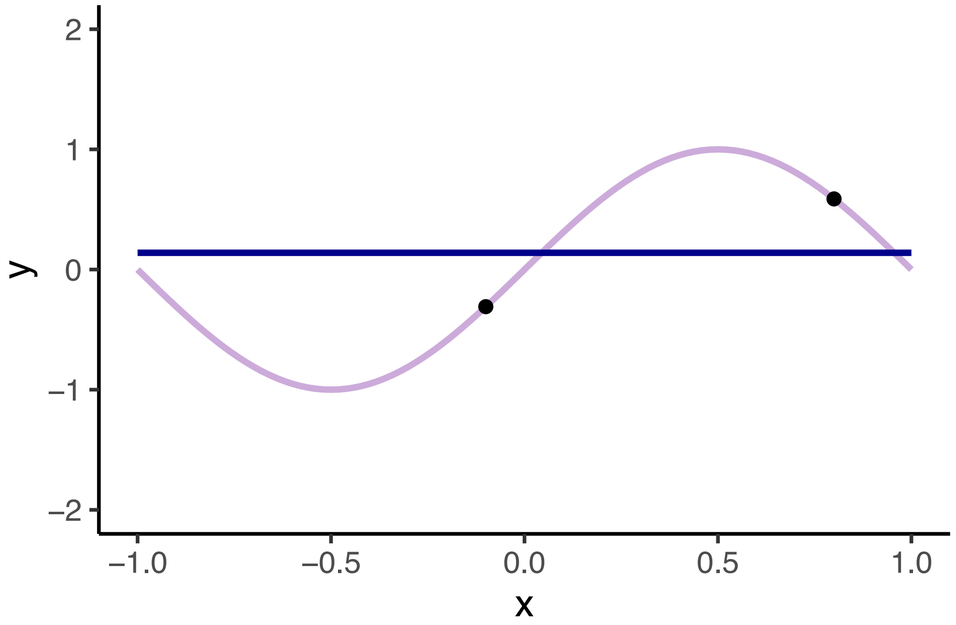
Duas classes de modelos

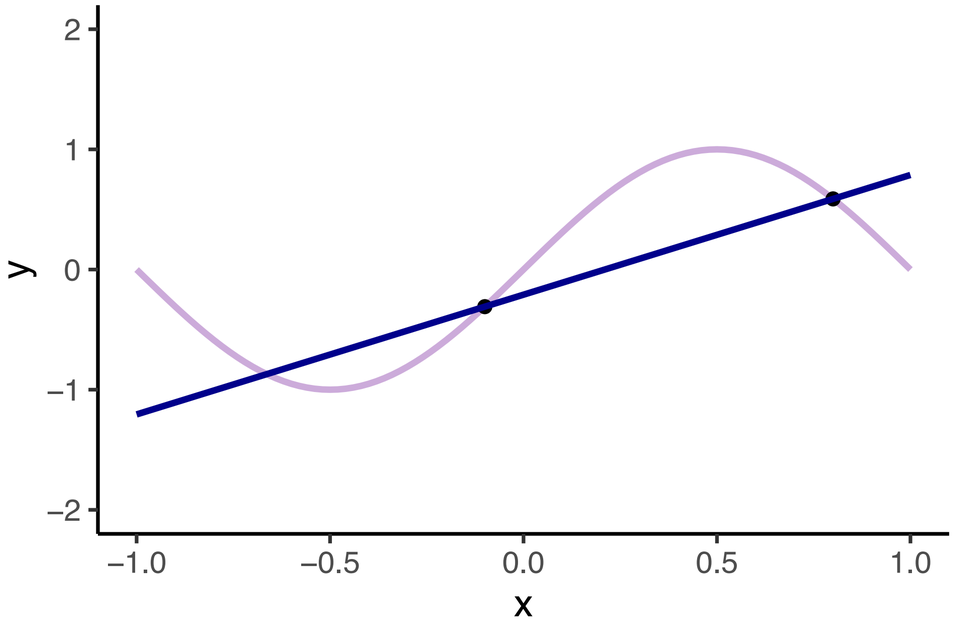
$\mathcal{G}_1 = \{g(x) = \beta_0 \colon \beta_0 \in \mathbb{R}\}$

$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x \colon (\beta_0, \beta_1) \in \mathbb{R}^2\}$



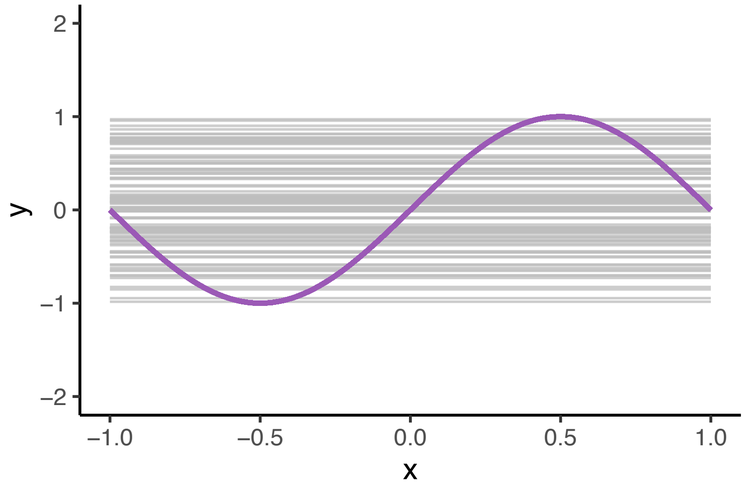
Ajuste versus complexidade

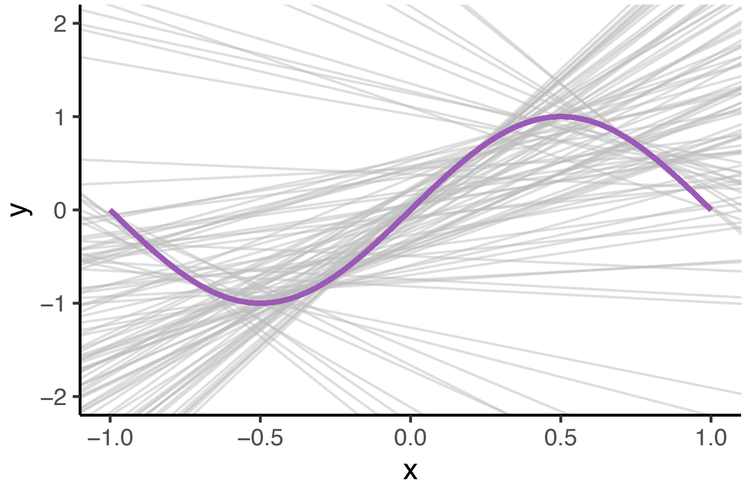


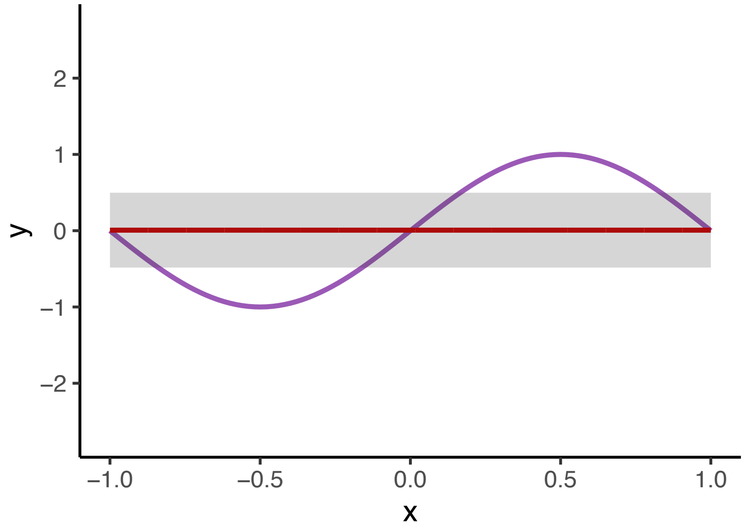


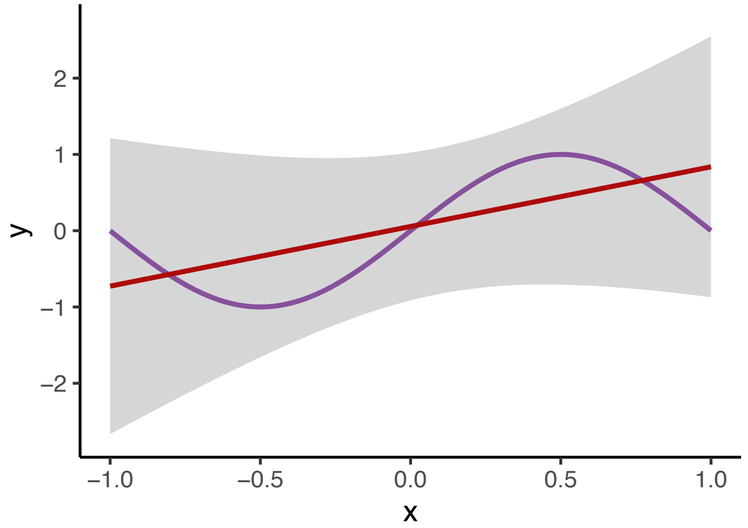
$$\mathcal{G}_1 = \{g(x) = \beta_0 \colon \beta_0 \in \mathbb{R}\}$$

$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x \colon (\beta_0, \beta_1) \in \mathbb{R}^2\}$$







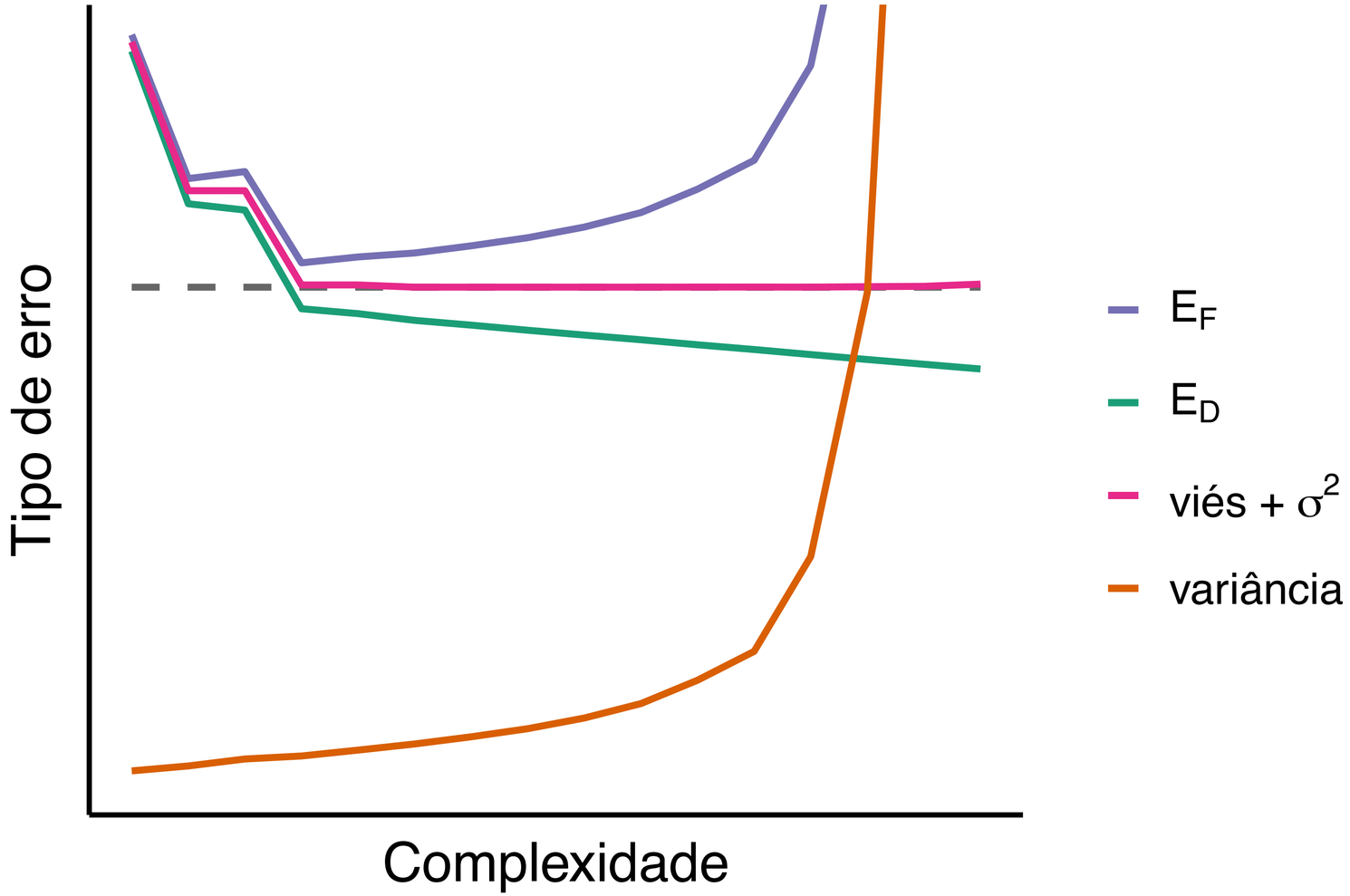


Ajuste versus complexidade

$\text{Variância} + \text{Viés}^2 = 0{,}75$

$\text{Variância} + \text{Viés}^2 = 1{,}90$

Curvas de erro



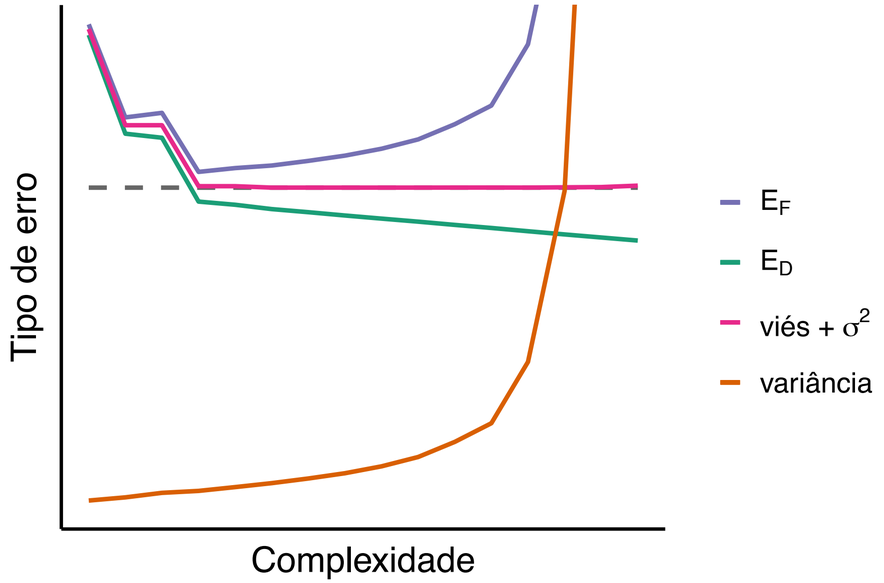
Lembrando o modelo linear para regressão

$$\mathcal{G} = \{g(x) = x^T\beta,\ \ \beta \in \mathbb{R}^{p+1}\}$$

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$x^T\beta = \beta_0 + \sum_{j=1}^{p} x_j\beta_j$$

$$\widehat{E}_D(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$



Escolhemos $\beta \in \mathbb{R}^{p+1}$ que minimiza $\widehat{E}_D(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$, denotamos por $\widehat{\beta}$ o vetor obtido e fazemos $g^{\mathcal{D}}(x) = x^T\widehat{\beta}$

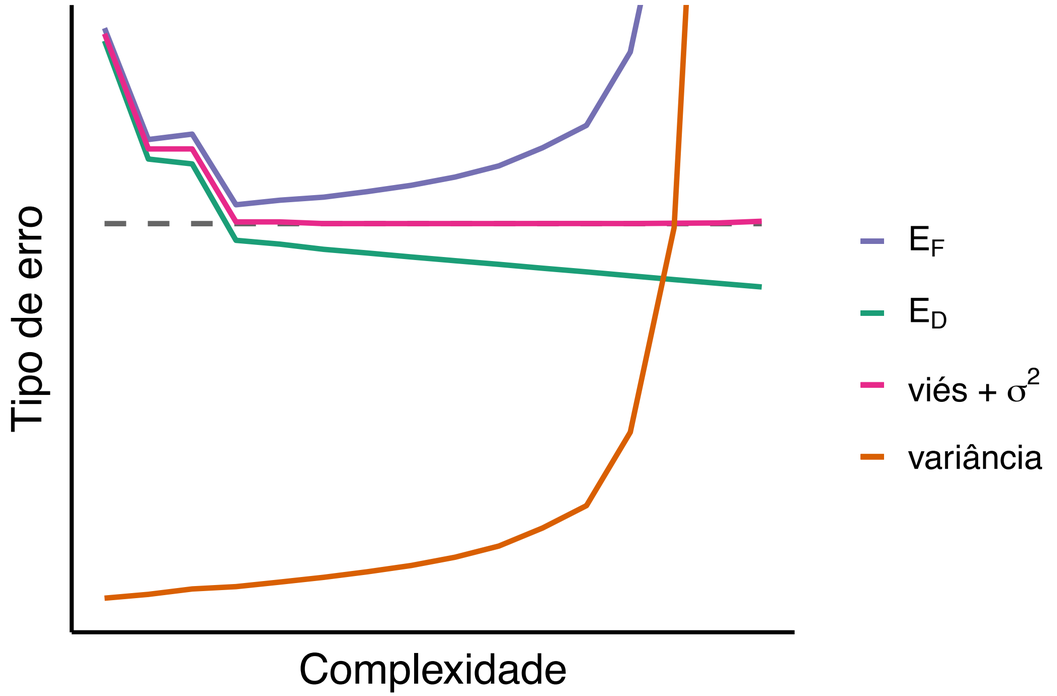
Seleção de variáveis

- No modelo linear em alta dimensão; isto é quando p é grande em relação a n, em geral não podemos usar o método de mínimos quadrados irrestrito
- Na maioria das vezes, estamos interessados em ajustar um modelo linear num *conjunto menor* de variáveis preditoras
- Esta seleção de variáveis tem a vantagem de aumentar a interpretabilidade do modelo e reduzir sua complexidade, evitando o superajuste
- Uma primeira ideia seria utilizar os métodos de validação/validação cruzada para escolher o melhor subconjunto de variáveis, mas este procedimento seria excessivamente demorado e não proporcionaria um único subconjunto de variáveis

Seleção de variáveis - BIC

- Algumas abordagens tentam aproximar com uma fórmula a curva de $E_F(g^{\mathscr{D}})$ em função da complexidade de $g^{\mathscr{D}}$, ajustando o erro estimado dentro da amostra com uma penalidade

- O critério mais conhecido para isso é o Critério Bayesiano da Informação (BIC), que está definido como $\text{BIC}(g) = \widehat{E}_D(g) + \frac{\log n}{n} d(g)\hat{\sigma}^2$, onde $d(g)$ representa a complexidade de g (número de variáveis) e $\hat{\sigma}^2$ é uma estimativa da variância do erro ϵ. Tipicamente, $\hat{\sigma}^2$ é estimado usando o maior modelo considerado

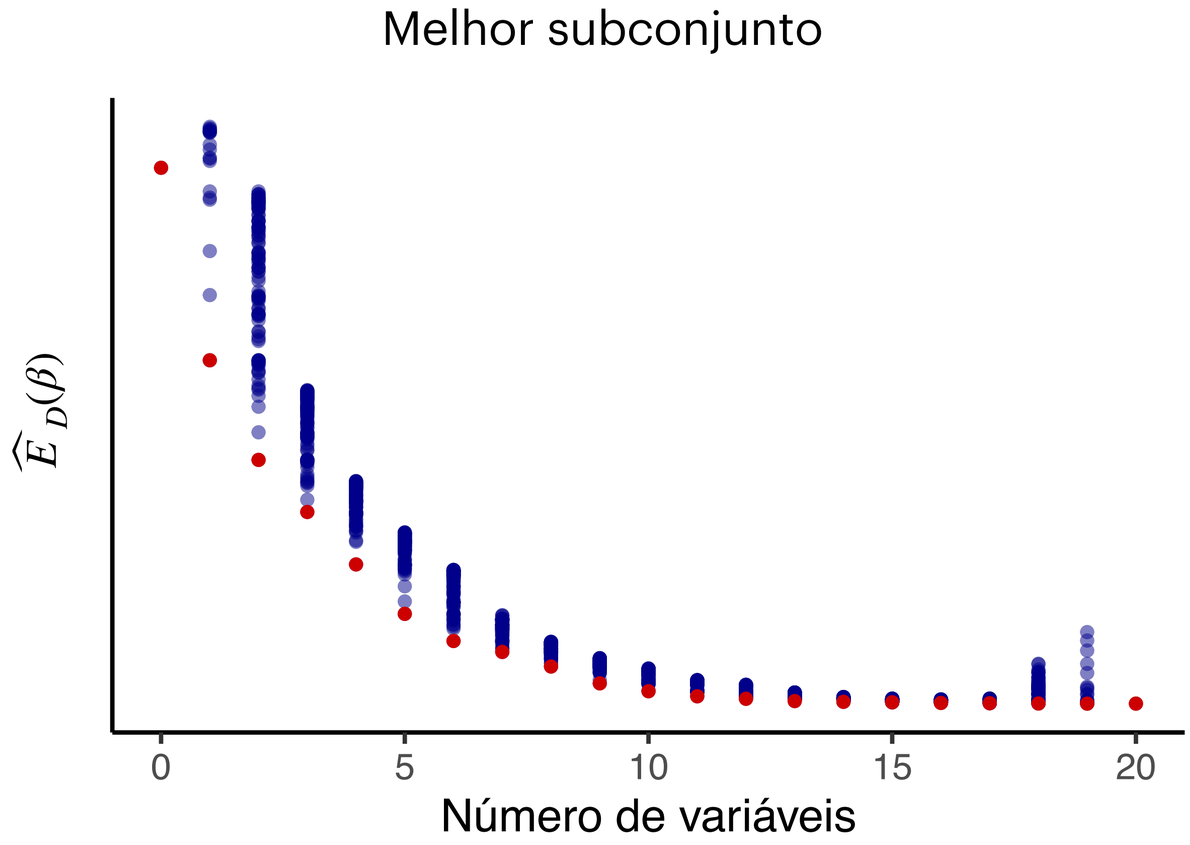


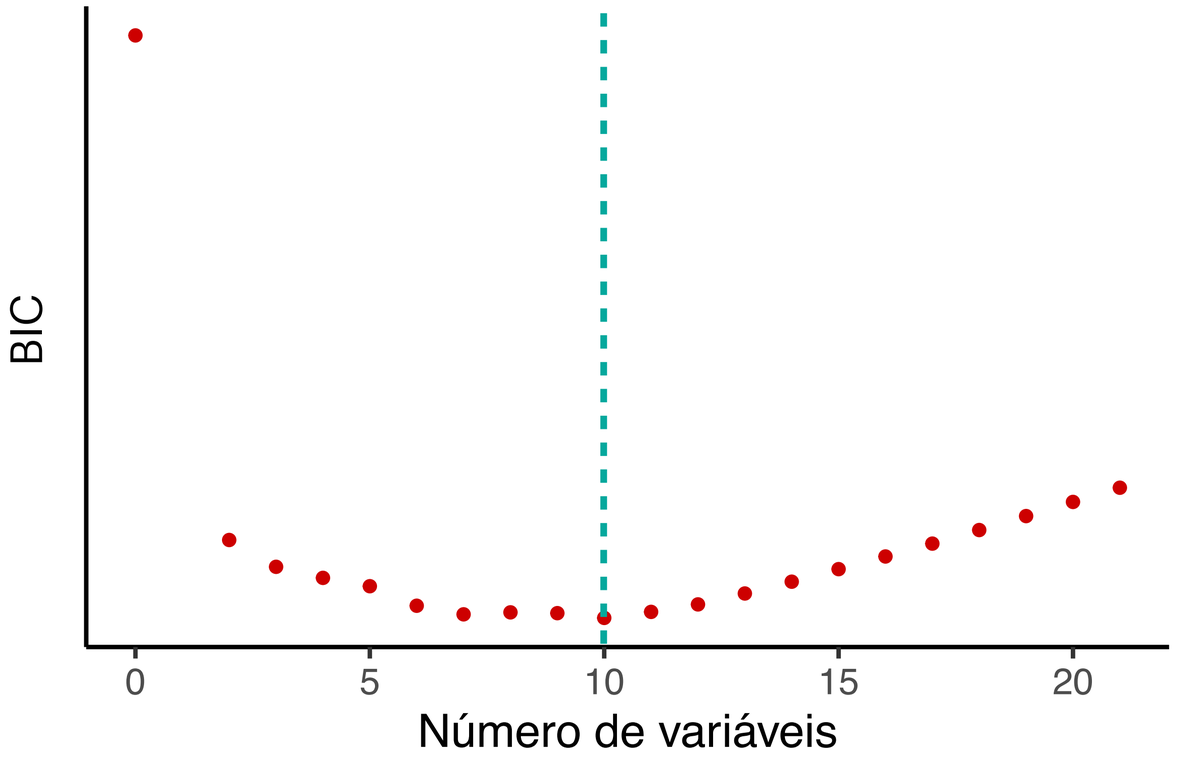
Seleção de variáveis - Melhor subconjunto

Algoritmo: melhor subconjunto

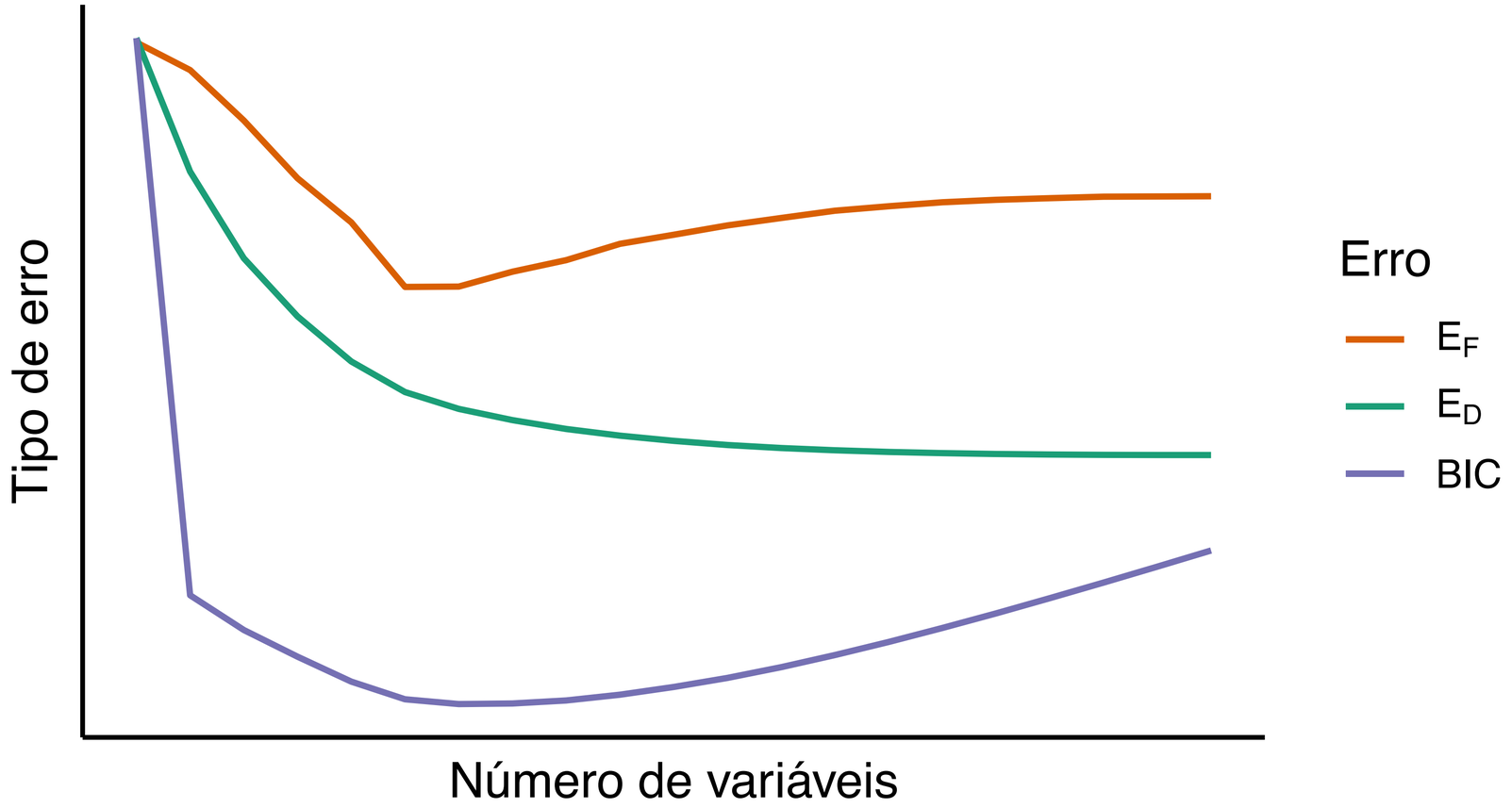
1. Seja $g_0(x) = \bar{y}$ para todo x. Este é o modelo *nulo*, que não contem nenhuma variável preditora.
2. Para $k = 1,2,\ldots,p$:
   (a) Ajuste todos os $\binom{p}{k}$ modelos que contêm exatamente k variáveis preditoras.
   (b) Escolha o modelo entre os $\binom{p}{k}$ possíveis com menor $\widehat{E}_D$ e chame a função correspondente de $g_k(x)$
3. Escolha a melhor função entre $g_0(x), \ldots, g_p(x)$ usando validação cruzada ou um método regularizado como BIC.

Seleção de variáveis - Melhor subconjunto





Seleção de variáveis - Melhor subconjunto



Seleção de variáveis - Melhor subconjunto

- Este algoritmo faz uma busca exaustiva no espaço de todos os subconjuntos: se p for grande isto é muito custoso e pode ser inviável

- Para fazer o algoritmo mais rápido e eficiente, pode-se restringir o número máximo de covariáveis; i.e obtendo somente as funções $g_0(x), \ldots, g_r(x)$ para algum $r < p$

Algoritmo: melhor subconjunto

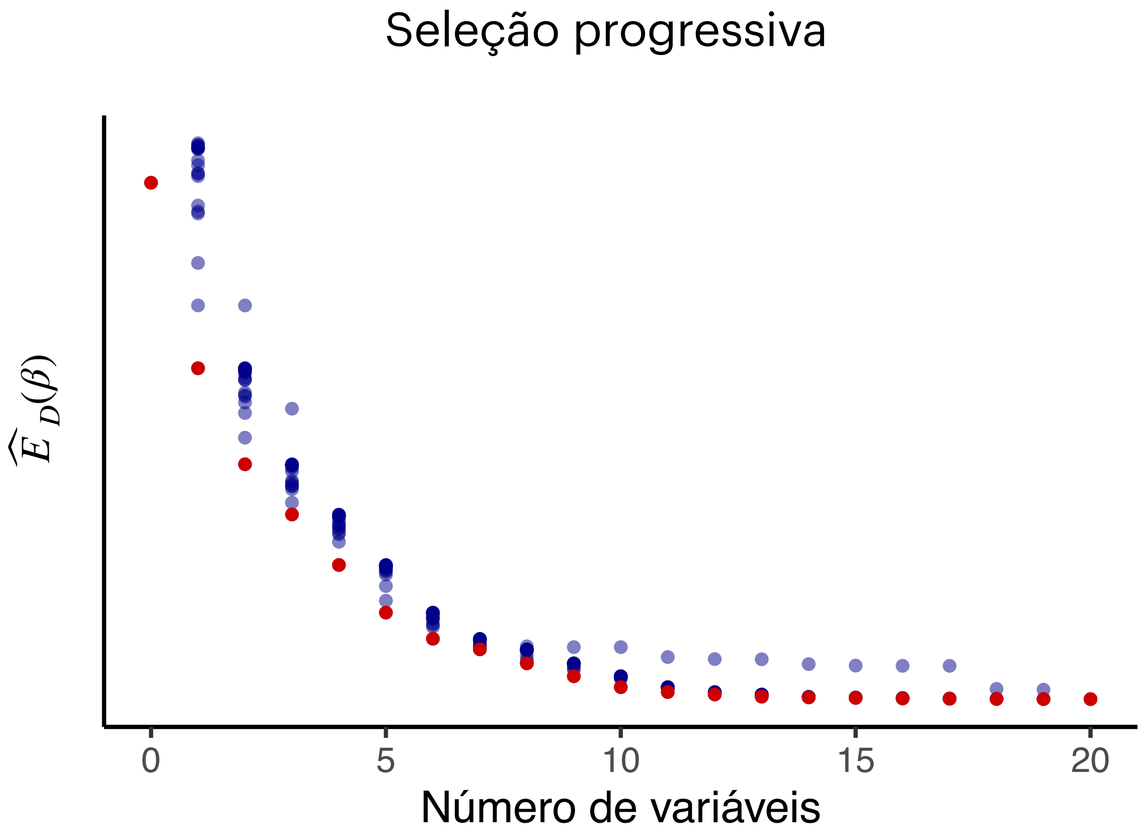
1. Seja $g_0(x) = \bar{y}$ para todo x. Este é o modelo *nulo*, que não contem nenhuma variável preditora.
2. Para $k = 1,2,\ldots,p$:
   (a) Ajuste todos os $\binom{p}{k}$ modelos que contêm exatamente k variáveis preditoras.
   (b) Escolha o modelo entre os $\binom{p}{k}$ possíveis com menor $\widehat{E}_D$ e chame a função correspondente de $g_k(x)$
3. Escolha a melhor função entre $g_0(x), \ldots, g_p(x)$ usando validação cruzada ou um método regularizado como BIC.

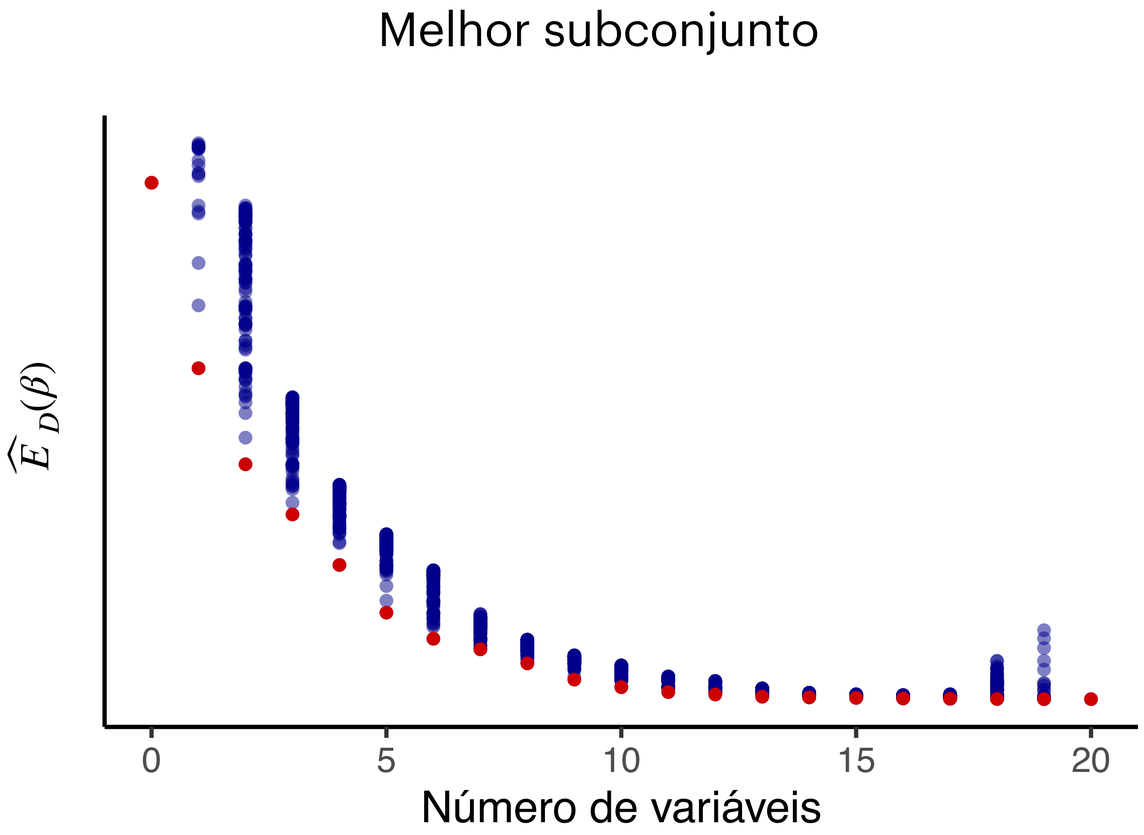
Seleção de variáveis - Seleção progressiva

Algoritmo: seleção progressiva (*forward stepwise selection*)

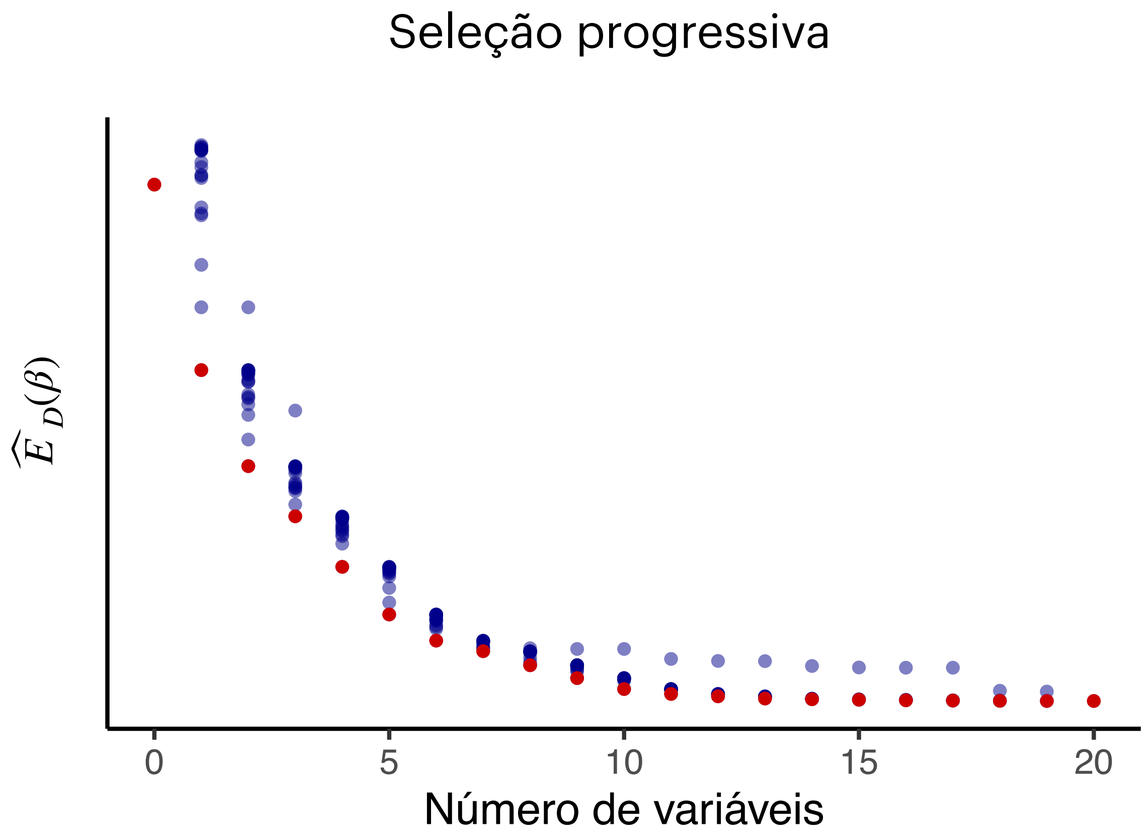
1. Seja $g_0(x) = \bar{y}$ para todo x. Este é o modelo *nulo*, que não contem nenhuma variável preditora.
2. Para $k = 0,2,\ldots,p-1$:
   (a) Considere todos os $p-k$ modelos que aumentam os preditores em $g_k(x)$ com um preditor adicional.
   (b) Escolha o modelo entre os $p-k$ possíveis com menor $\widehat{E}_D$ e denote-o por $g_{k+1}(x)$.
3. Escolha a melhor função entre $g_0(x), \ldots, g_p(x)$ usando validação cruzada ou um método regularizado como BIC.

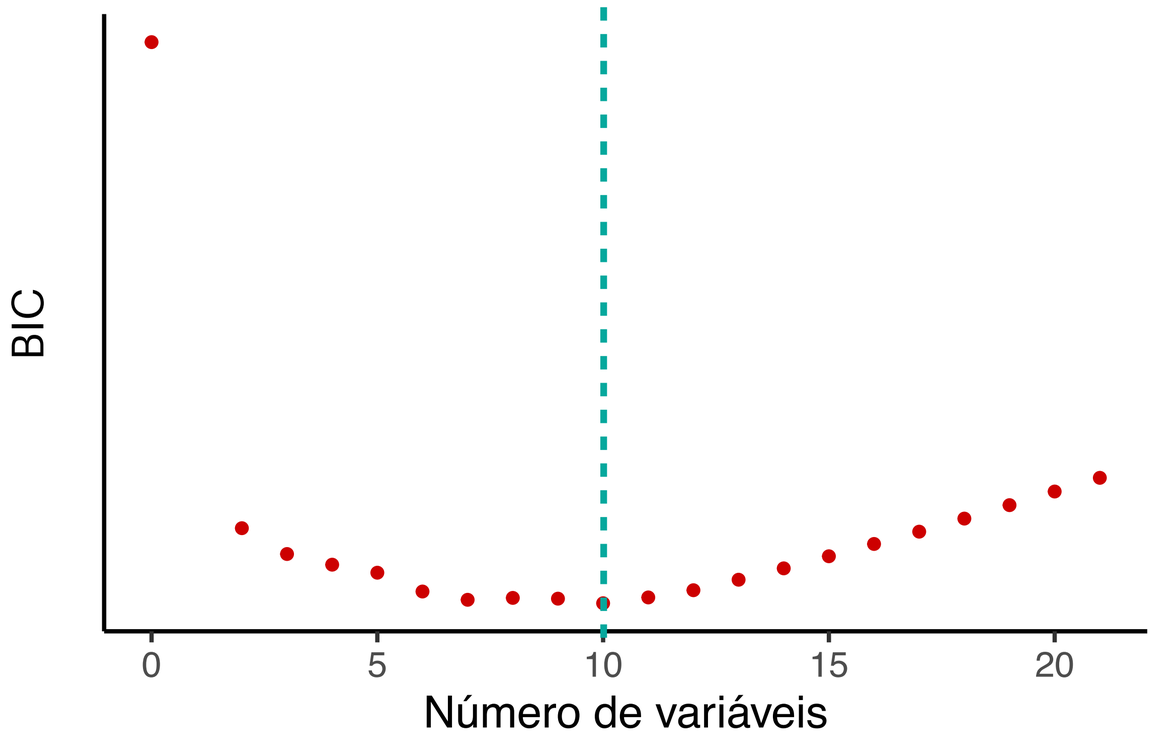
Seleção de variáveis - Seleção progressiva





Seleção de variáveis - Seleção progressiva



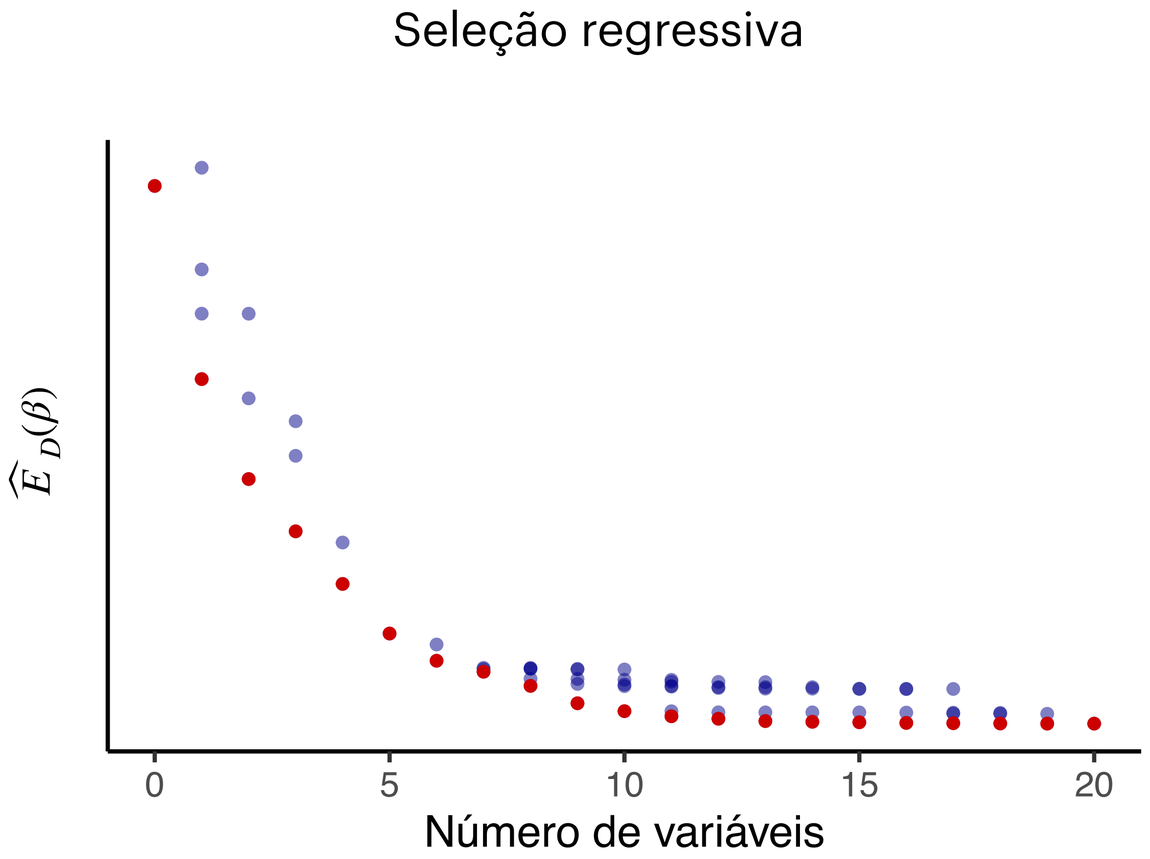


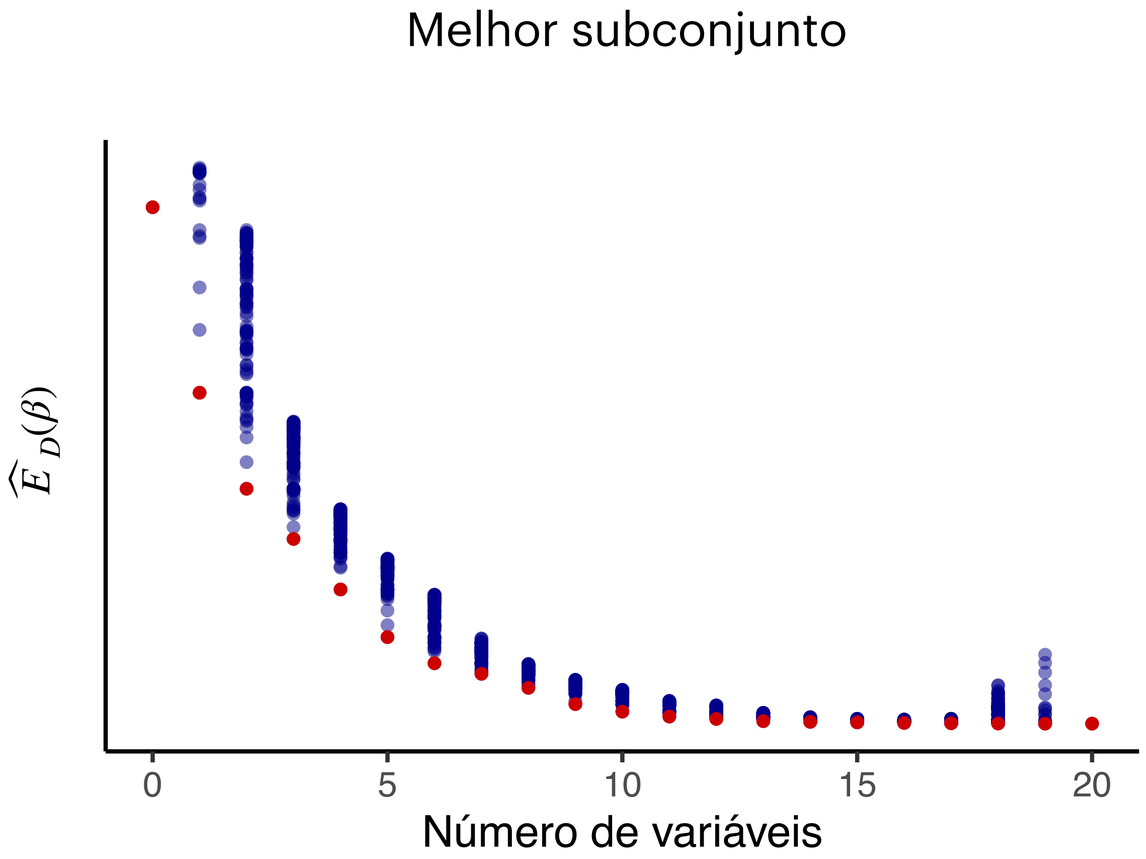
Seleção de variáveis - Seleção regressiva

Algoritmo: seleção regressiva (*backward stepwise selection*)

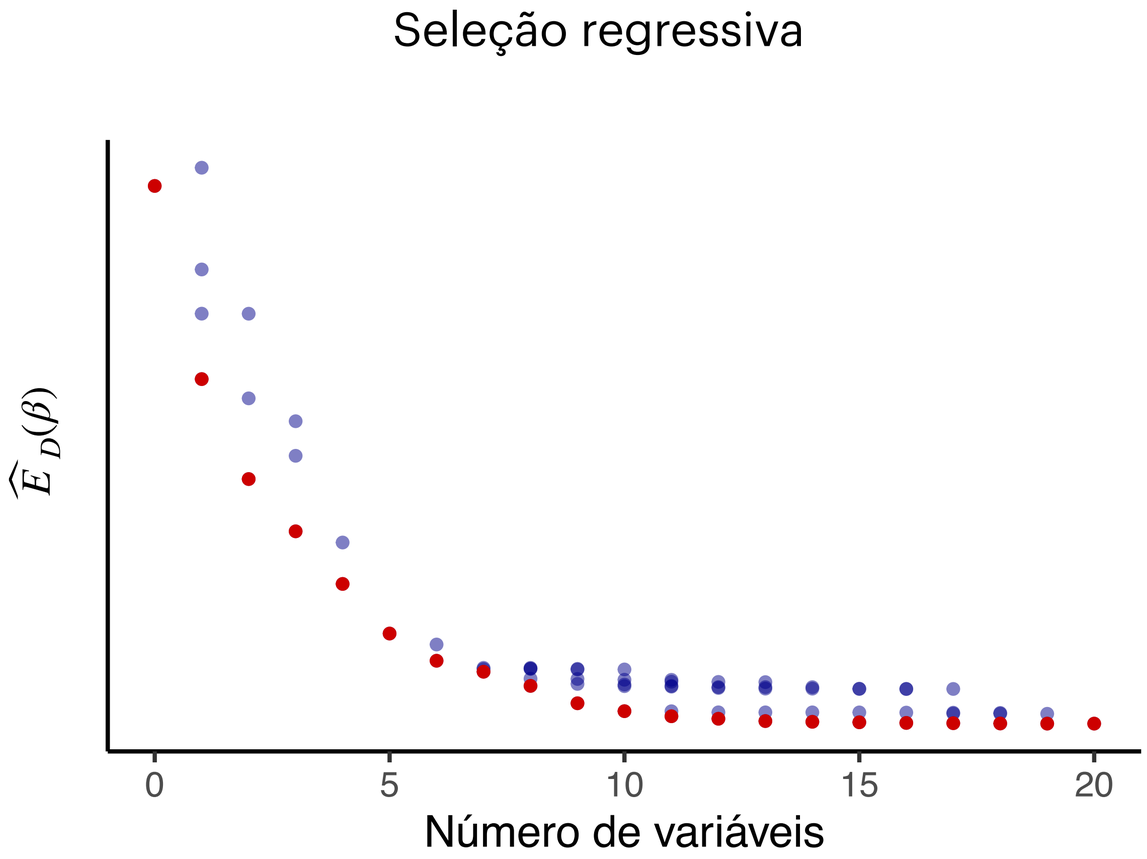
1. Seja $g_p(x)$ o maior modelo possível, aquele que contem todas as variáveis preditoras.
2. Para $k = p, p-1, \ldots, 1$:
   (a) Considere todos os k modelos que diminuem o número de variáveis preditoras em $g_k(x)$ retirando uma variável por vez.
   (b) Escolha o modelo entre os k possíveis com menor $\widehat{E}_D$ e denote-o por $g_{k-1}(x)$.
3. Escolha a melhor função entre $g_0(x), \ldots, g_p(x)$ usando validação cruzada ou um método regularizado como BIC.

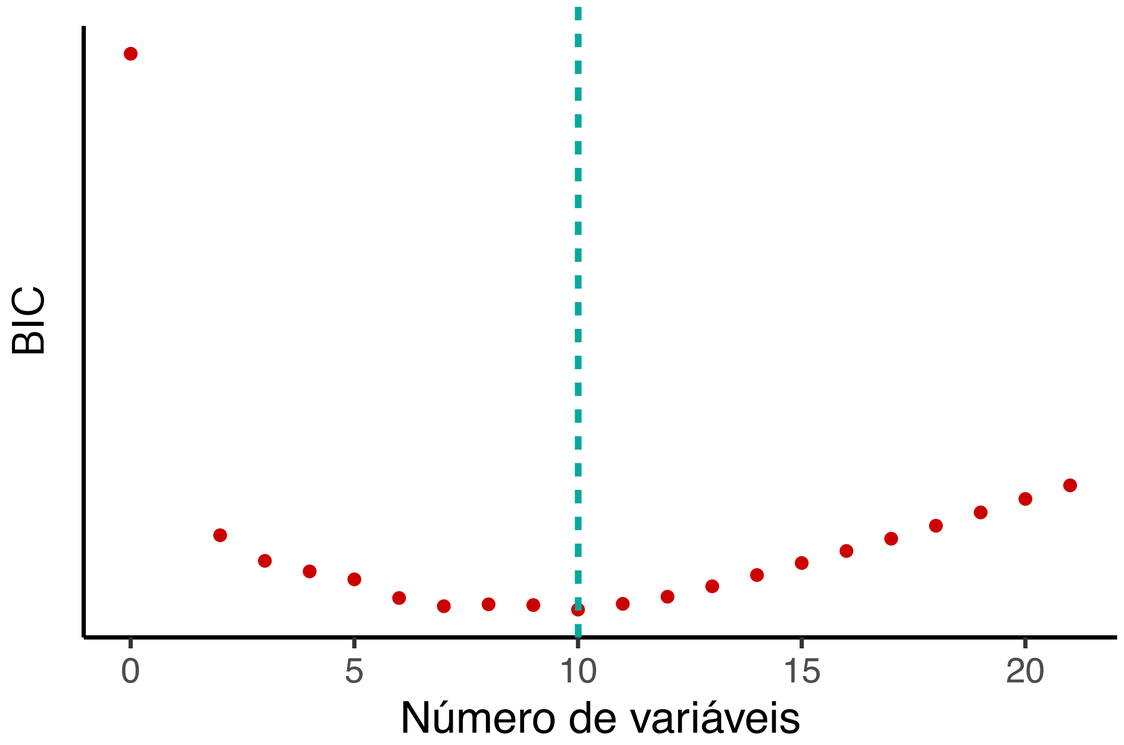
Seleção de variáveis - Seleção regressiva



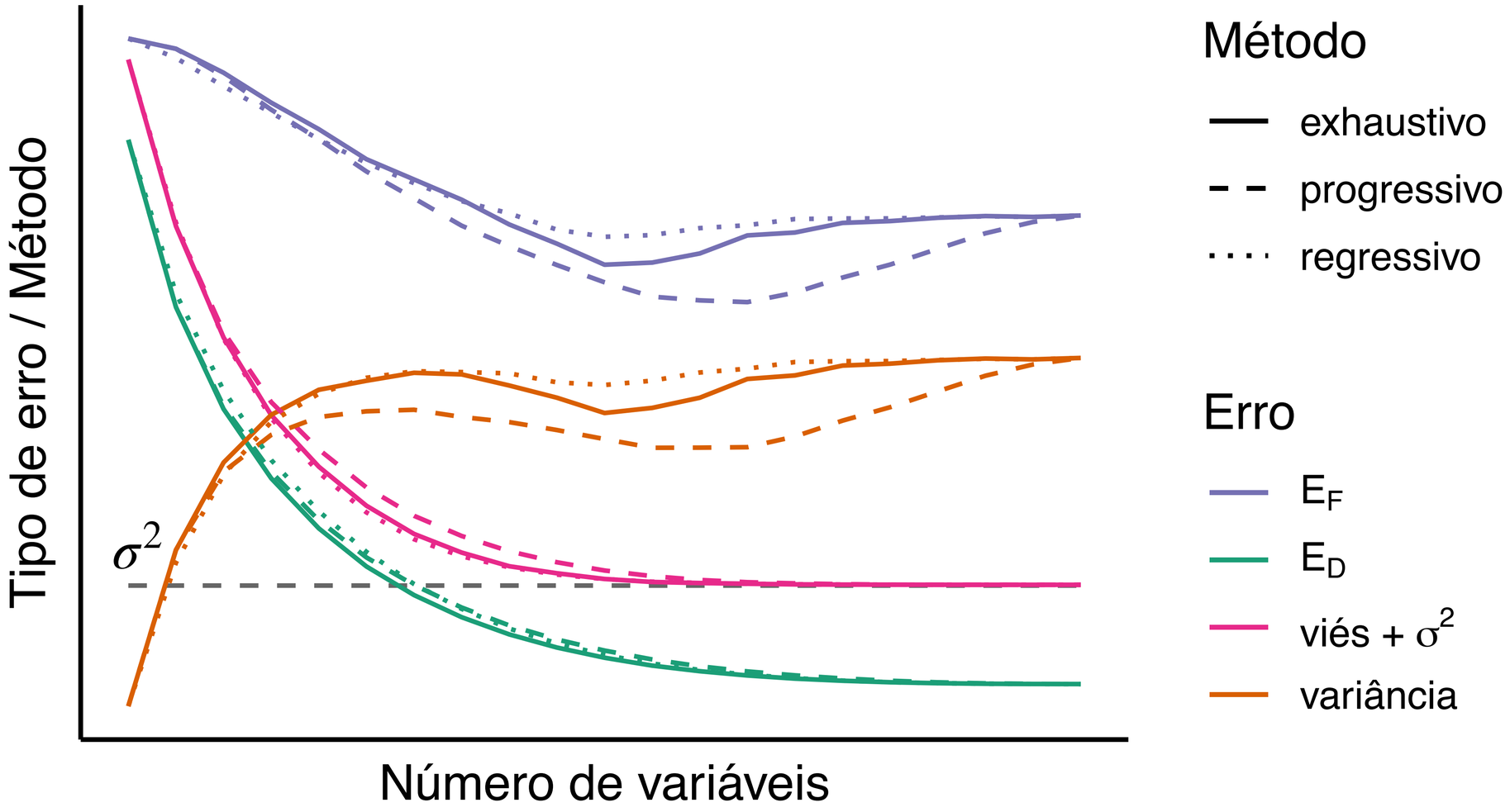


Seleção de variáveis - Seleção regressiva

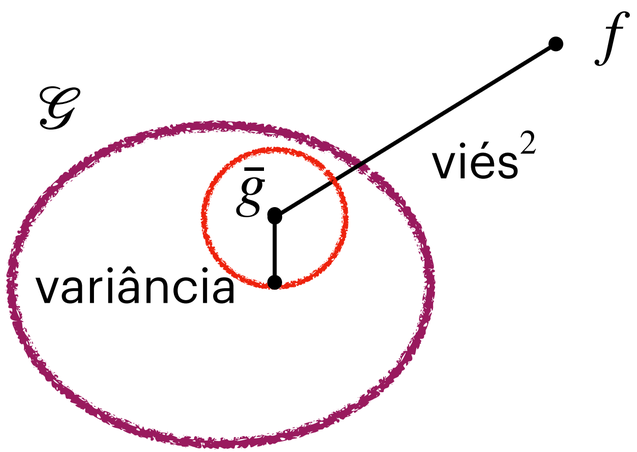




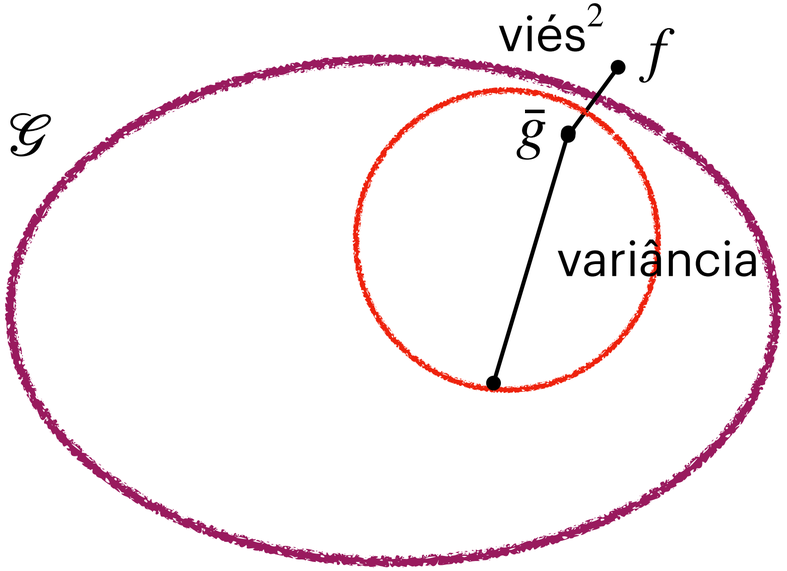
Curvas de erro



Diferentes modelos - Mesmo método/algoritmo

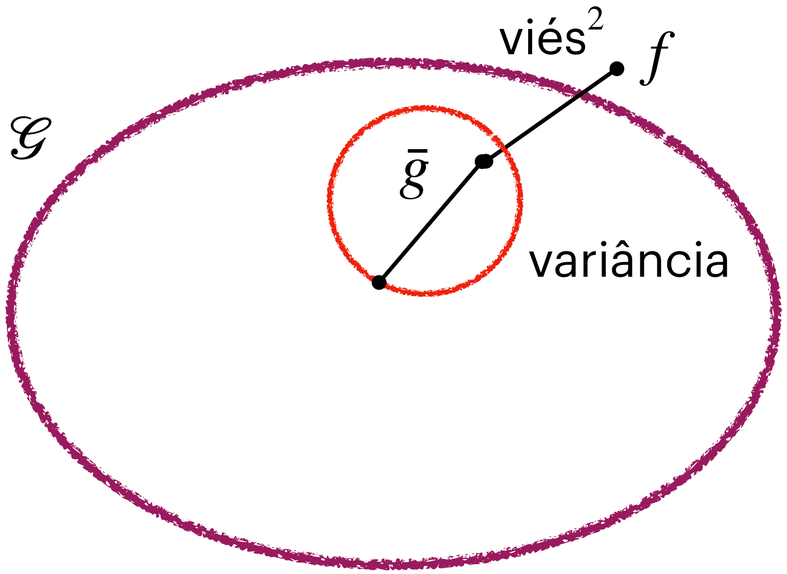


Modelo “simples”

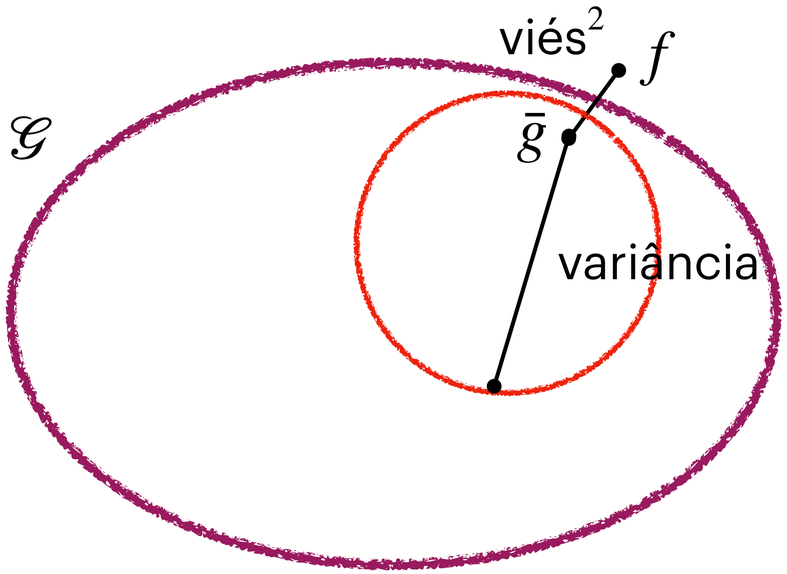


Modelo “complexo”

Mesmo modelo - Diferente método/algoritmo



Algoritmo “guloso”



Algoritmo “exhaustivo”

Conjunto de funções restrito - Regularização

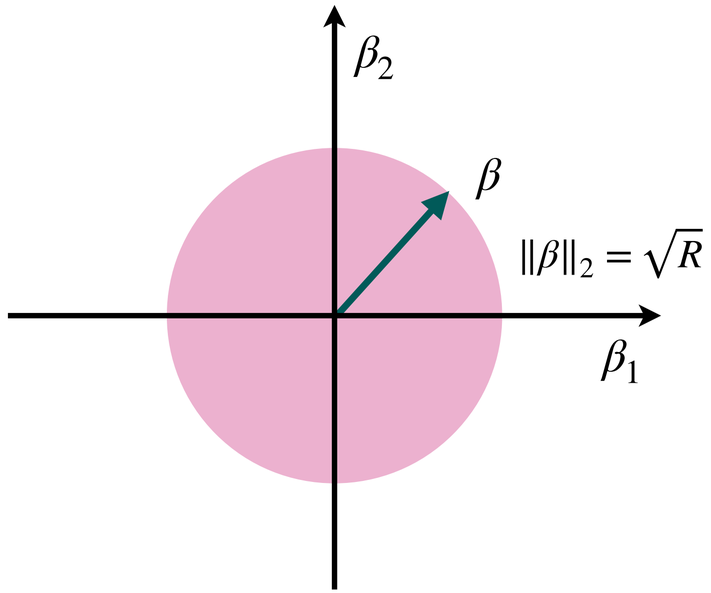
Num modelo “restrito”, em vez de deixarmos os valores de β (sem intercepto) variar livremente em $\mathbb{R}^p$, adicionamos uma restrição no conjunto de funções

$$\mathcal{G}_{ridge} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \|\beta\|_2^2 \leq R\}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\|\beta\|_2 = \sqrt{\sum_{1=1}^{p} |\beta_i|^2}$$

$$\|\beta\|_2^2 = \sum_{1=1}^{p} |\beta_i|^2$$



Mínimos quadrados regularizados - RIDGE

$$\mathcal{G}_{ridge} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \|\beta\|_2^2 \leq R\}$$

Neste modelo, queremos escolher:

$$g \in \mathcal{G}_{ridge} \quad \text{que minimize} \quad \widehat{E}_D(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2$$

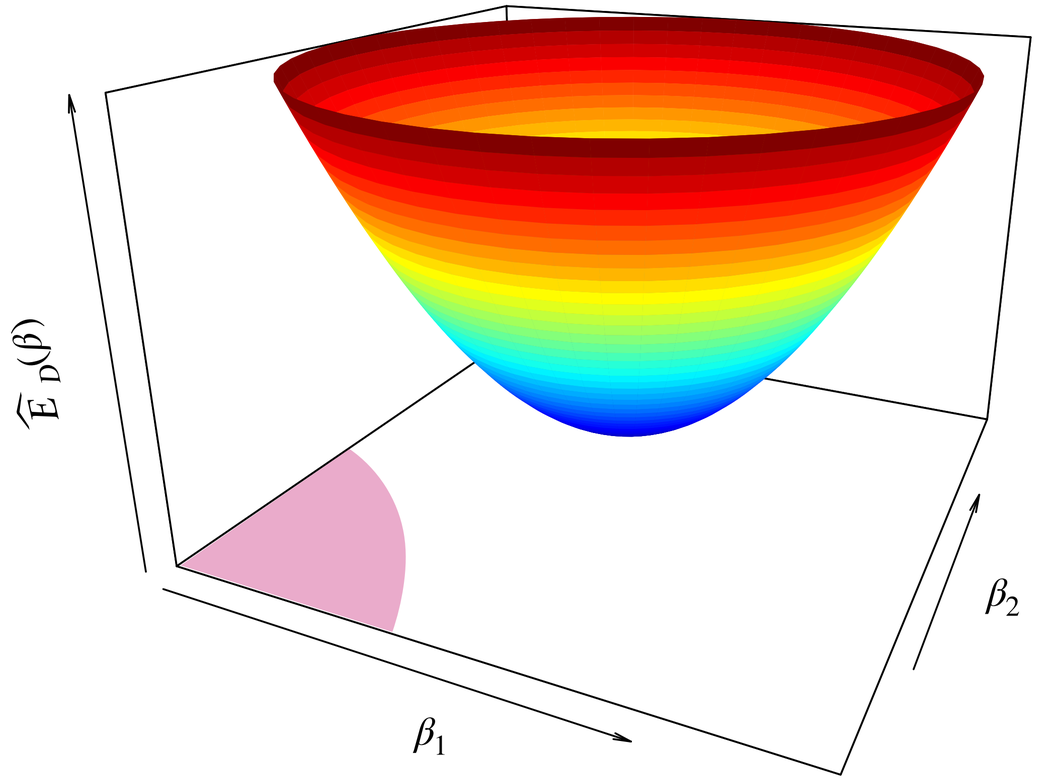
Isto é equivalente a escolher:

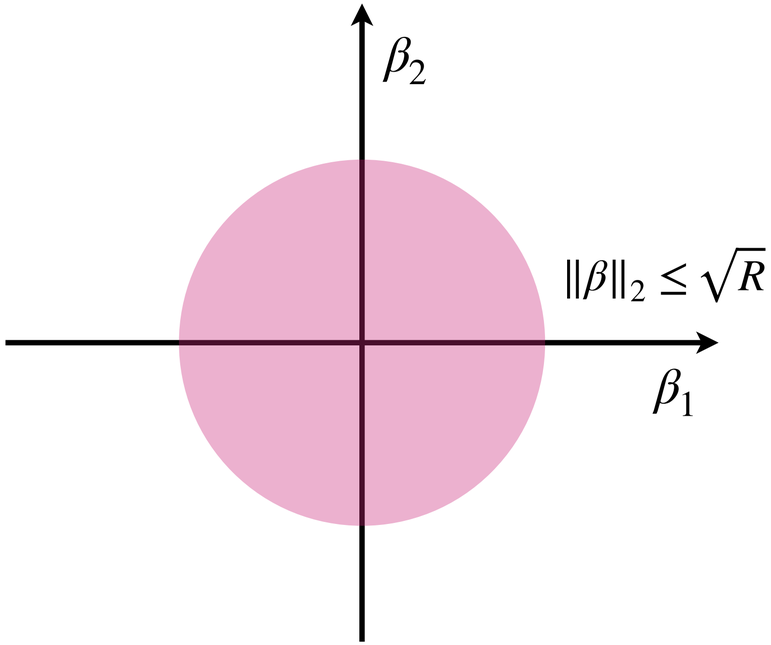
$$\beta \in \mathbb{R}^{p+1} \quad \text{que minimize} \quad \widehat{E}_D^{ridge}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_2^2$$

onde $\lambda \geq 0$ depende de R e da amostra $\mathcal{D}$

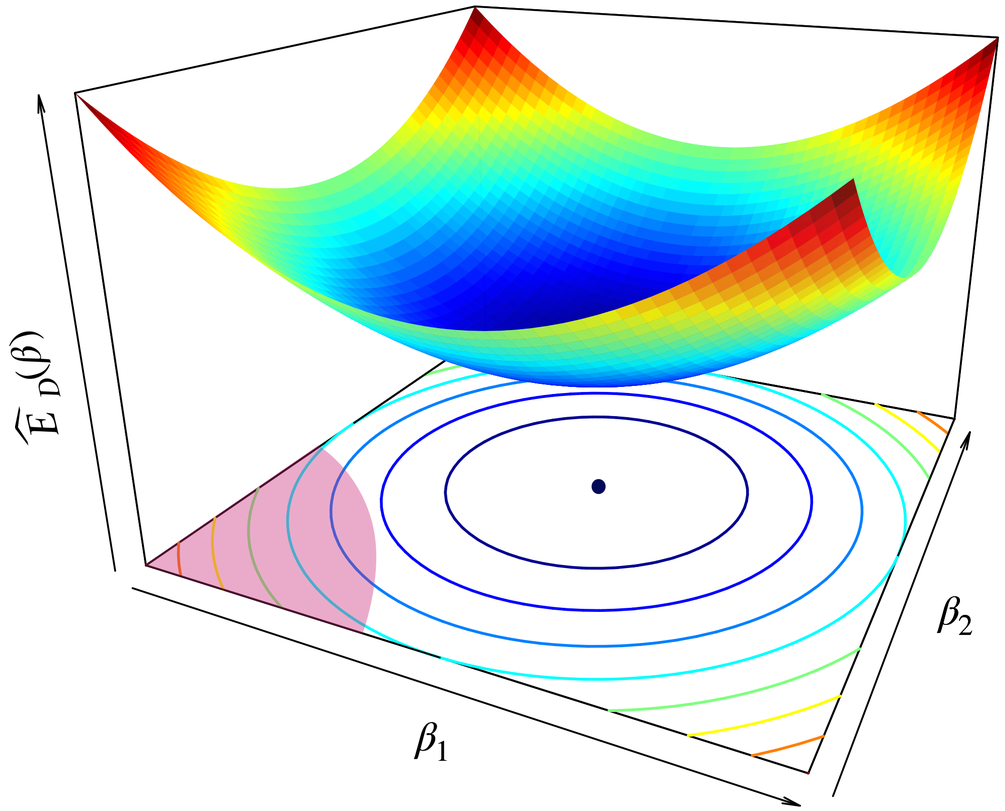
Este método é conhecido como “regressão RIDGE”

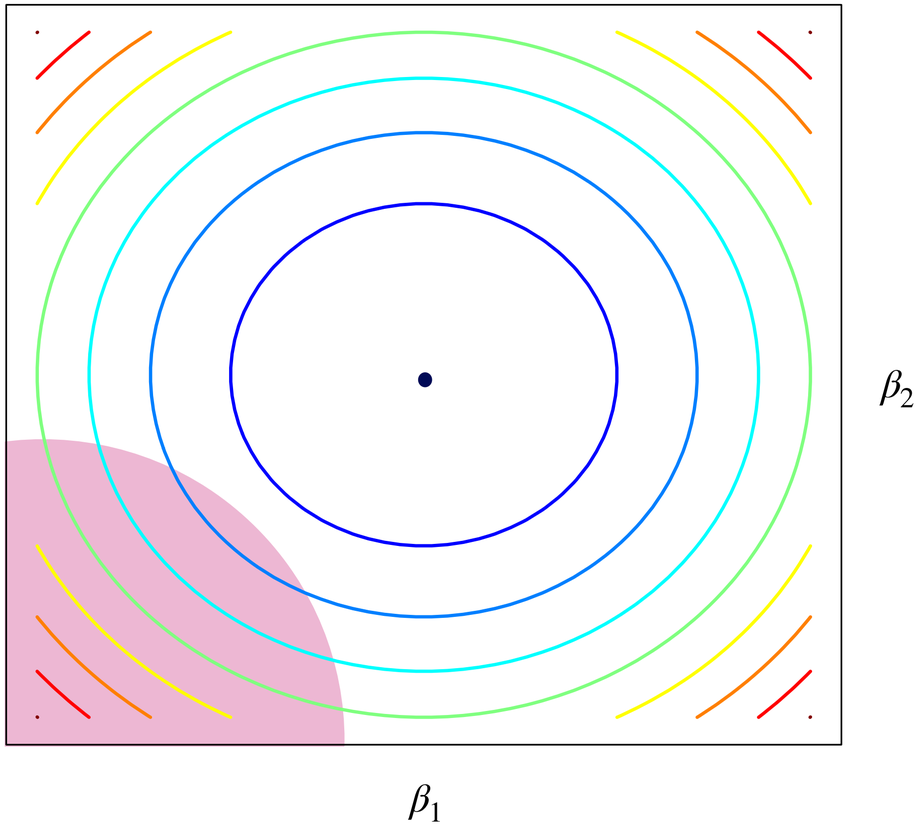
Mínimos quadrados regularizados - RIDGE



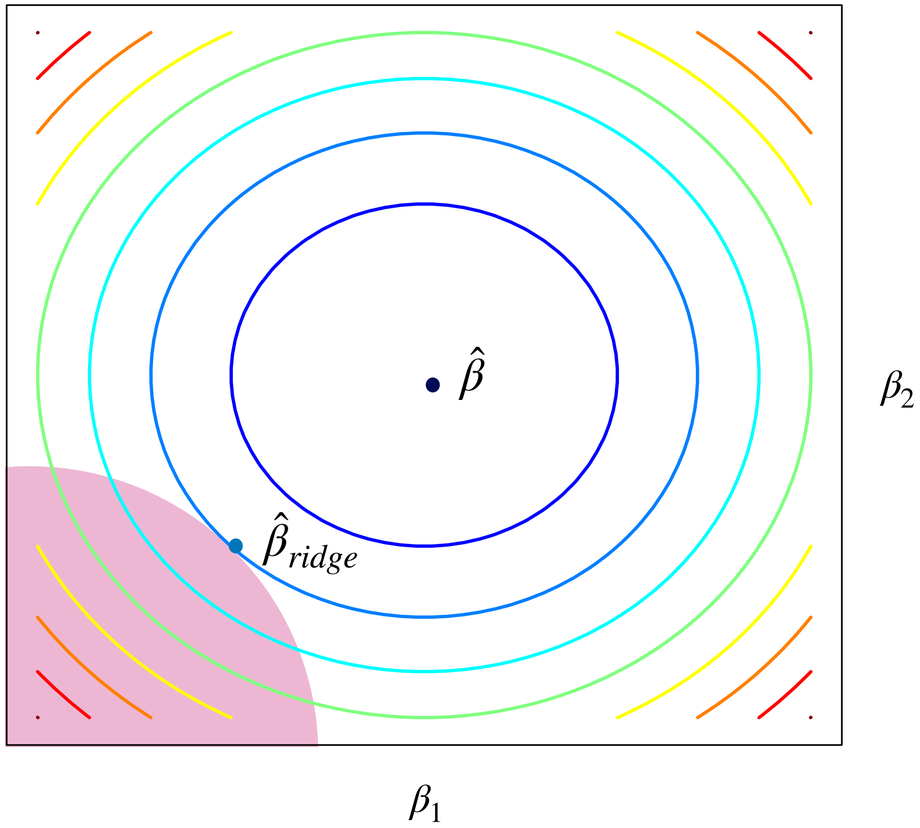


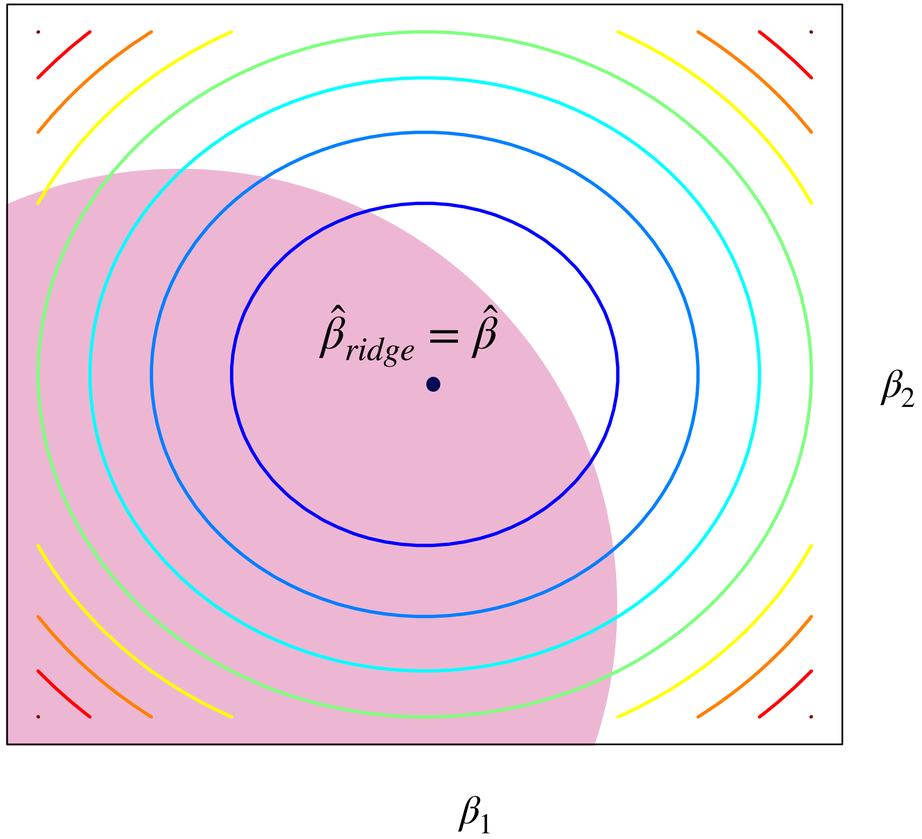
Mínimos quadrados regularizados - RIDGE





Mínimos quadrados regularizados - RIDGE





Mínimos quadrados regularizados - RIDGE

$$\mathcal{G}_{ridge} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \|\beta\|_2^2 \leq R\}$$

Queremos escolher $\beta \in \mathbb{R}^{p+1}$ que minimize $\widehat{E}_D^{ridge}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_2^2$

onde $\lambda \geq 0$ depende de R e da amostra $\mathcal{D}$

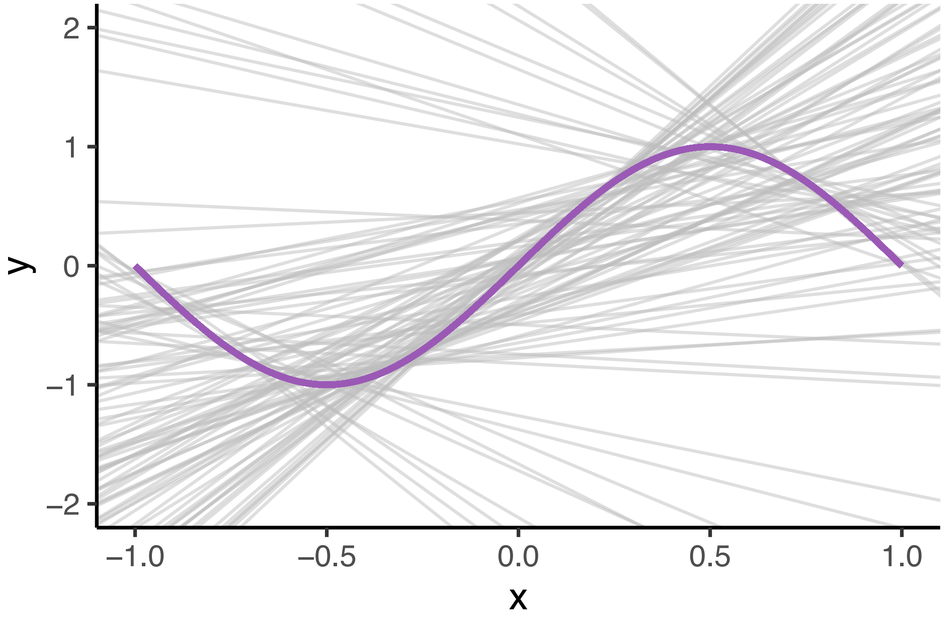
Solução de mínimos quadrados → $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

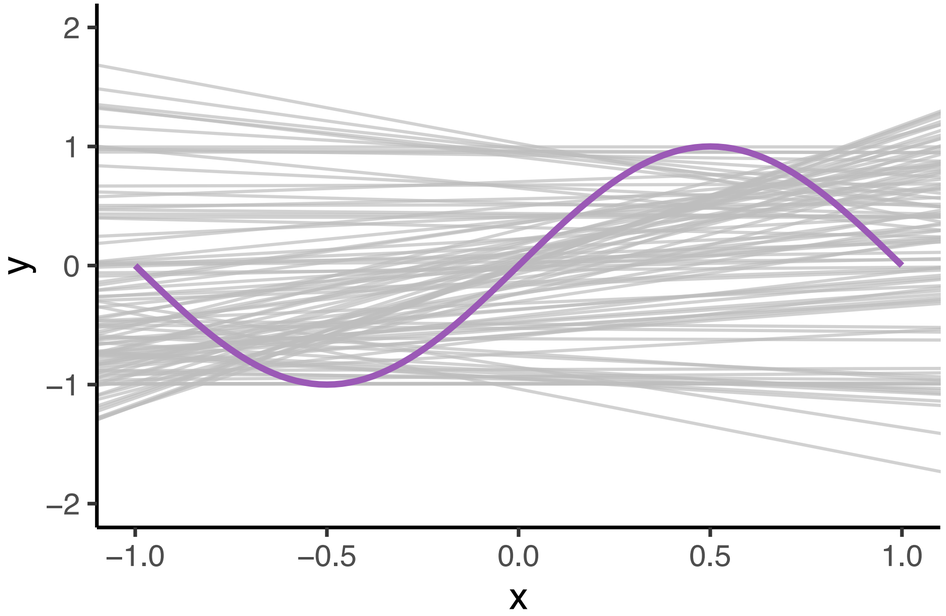
Solução de mínimos quadrados com penalidade RIDGE → $\hat{\beta}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

A inversa de $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p$ sempre existe se $\lambda > 0$

Exemplo

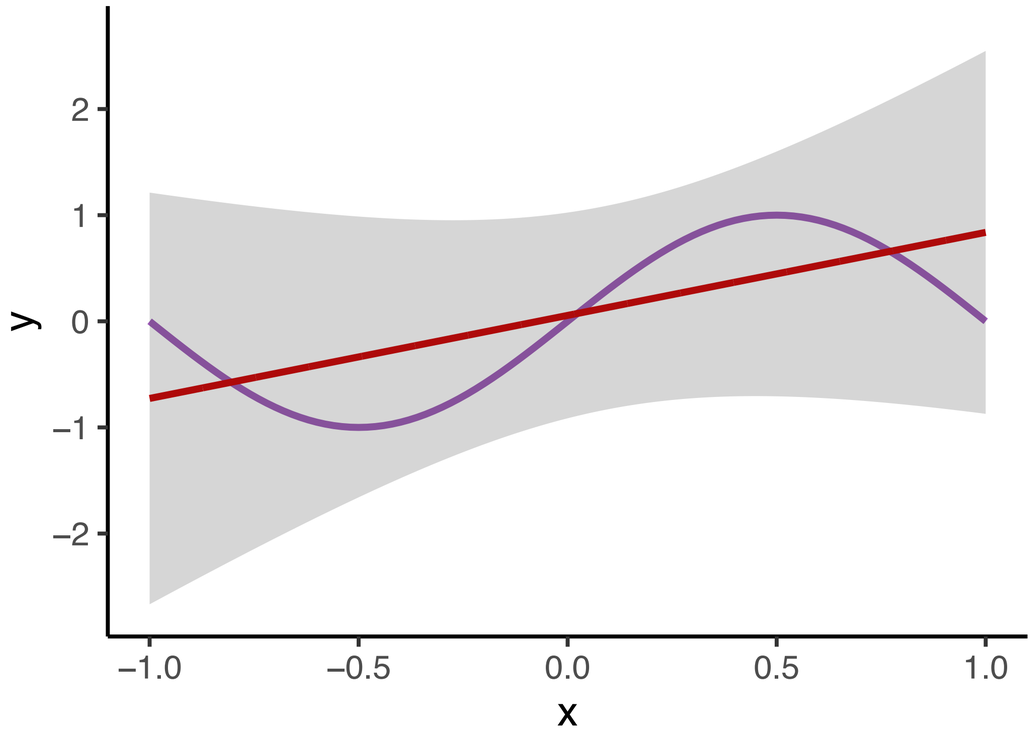




$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x \colon (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

$$\mathcal{G}_{ridge} = \{g(x) = \beta_0 + \beta_1 x\,,\ |\beta_1|^2 \leq R\}$$

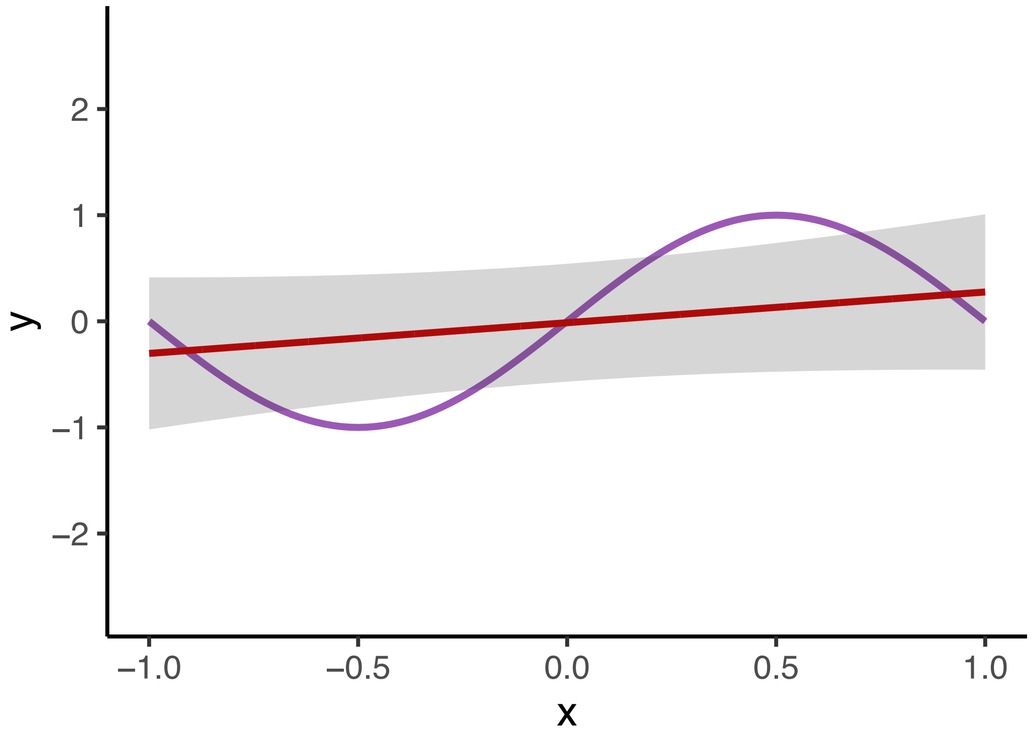
Decomposição do erro em viés e variância



$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x\,, (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

$$\text{Viés}^2 = 0{,}21$$

$$\text{Variância} = 0{,}69$$



$$\mathcal{G}_{ridge} = \{g(x) = \beta_0 + \beta_1 x\,,\ \ |\beta_1|^2 \leq R\}$$

$$\text{Viés}^2 = 0{,}34$$

$$\text{Variância} = 0{,}38$$

Mínimos quadrados regularizados - LASSO

$$\mathcal{G}_{lasso} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \|\beta\|_1 \leq R\}$$

Neste modelo, queremos escolher:

$$g \in \mathcal{G}_{lasso} \quad \text{que minimize} \quad \widehat{E}_D(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2$$

Isto é equivalente a escolher:

$$\beta \in \mathbb{R}^{p+1} \quad \text{que minimize} \quad \widehat{E}_D^{lasso}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1$$

onde $\lambda \geq 0$ depende de R e da amostra $\mathcal{D}$

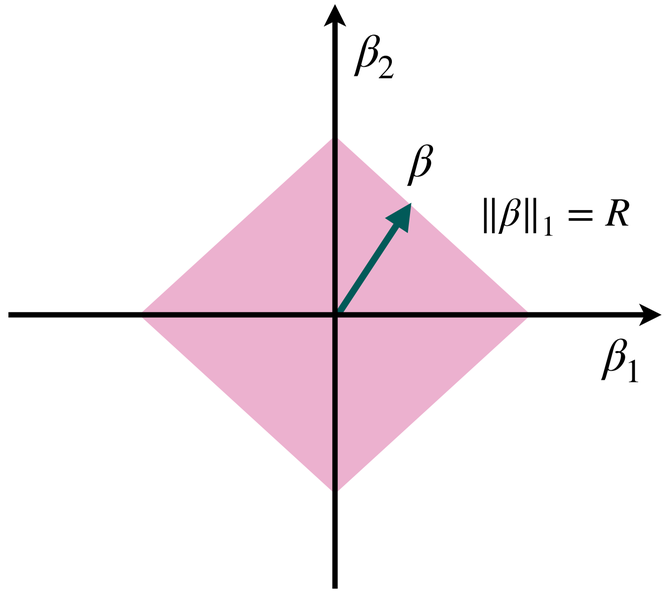
Este método é conhecido como “regressão LASSO”

Mínimos quadrados regularizados - LASSO

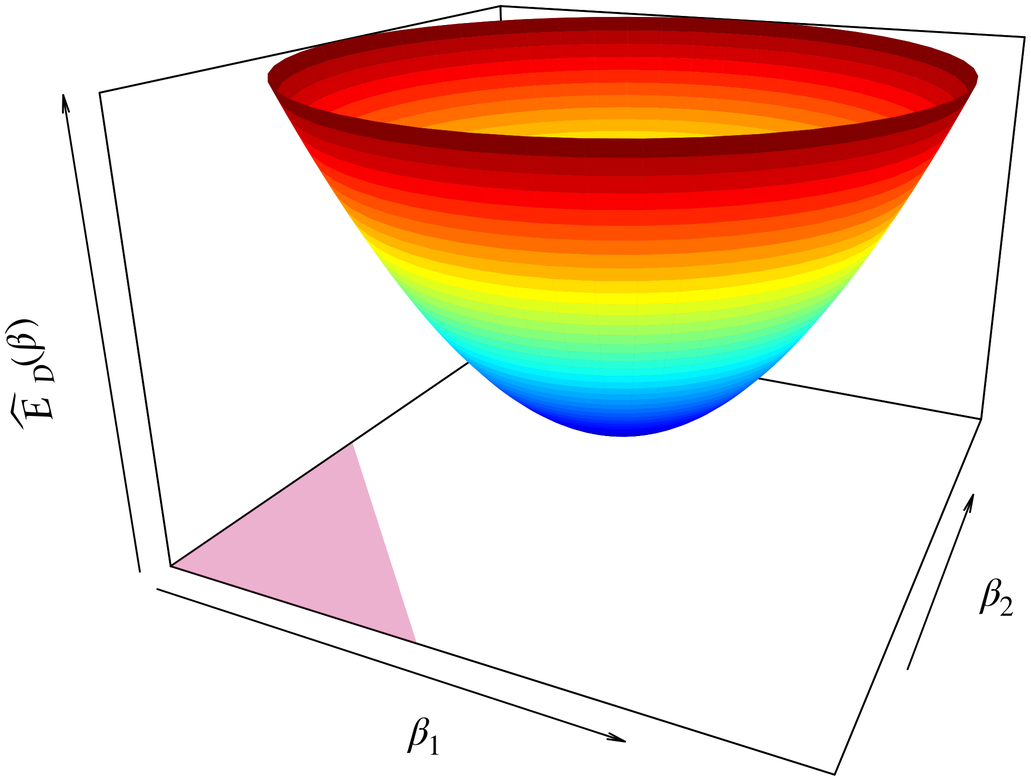
$$\mathcal{G}_{lasso} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p,\ \|\beta\|_1 \leq R\}$$

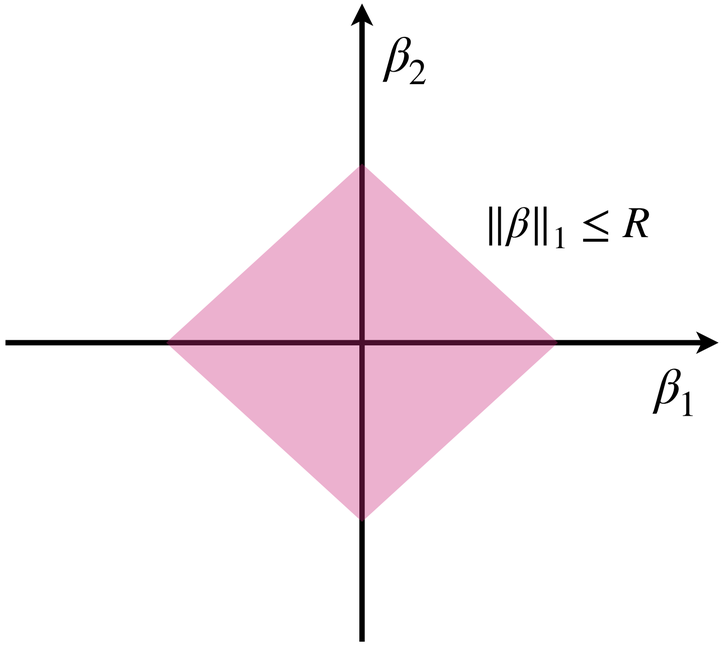
$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\|\beta\|_1 = \sum_{1=1}^{p} |\beta_i|$$

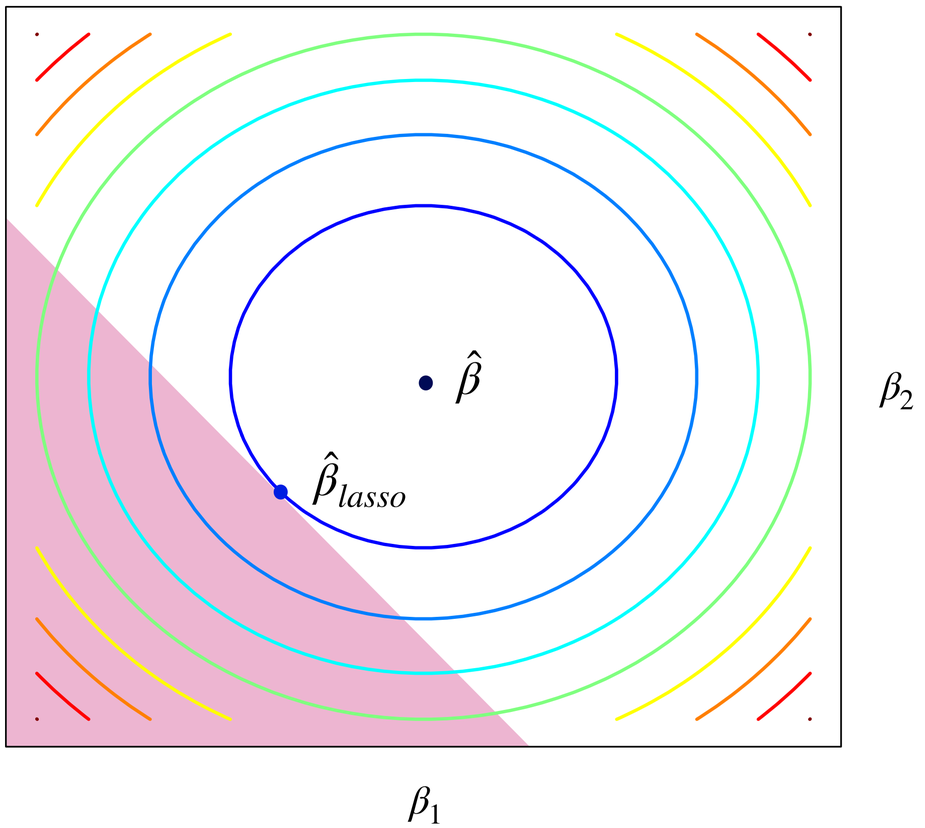


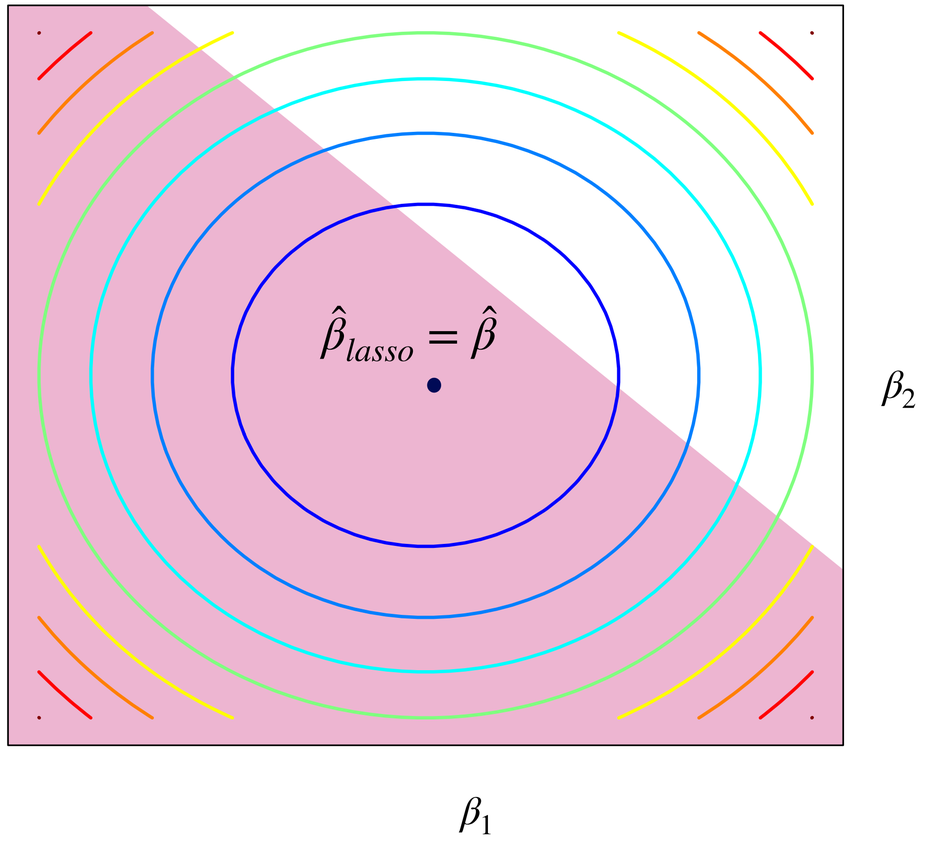
Mínimos quadrados regularizados - LASSO





Mínimos quadrados regularizados - LASSO





Mínimos quadrados regularizados - LASSO

$$\mathcal{G}_{lasso} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p,\ \|\beta\|_1 \leq R\}$$

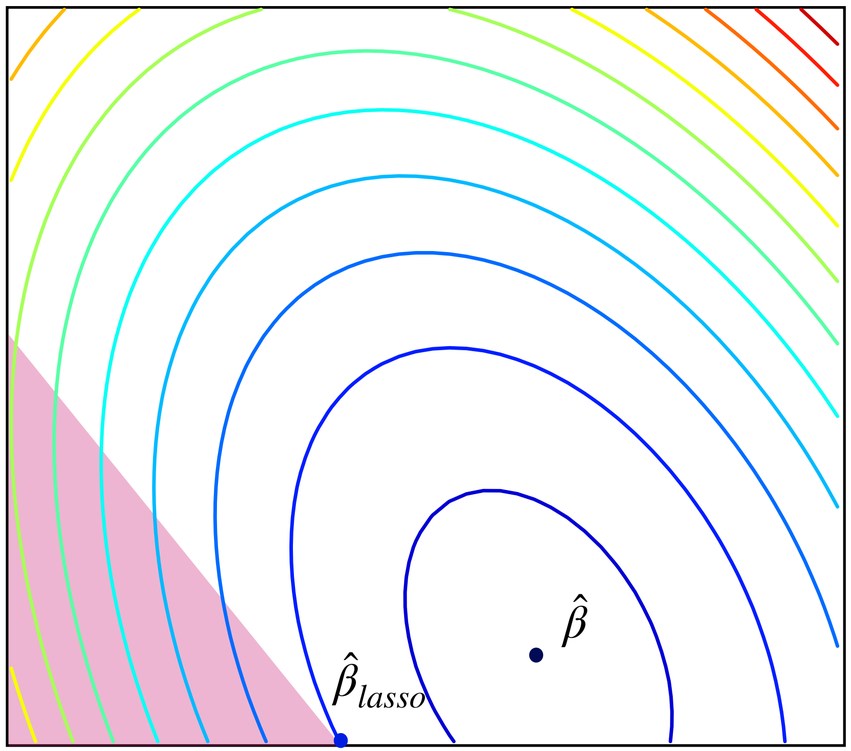
Queremos escolher $\beta \in \mathbb{R}^{p+1}$ que minimize $\widehat{E}_D^{lasso}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1$

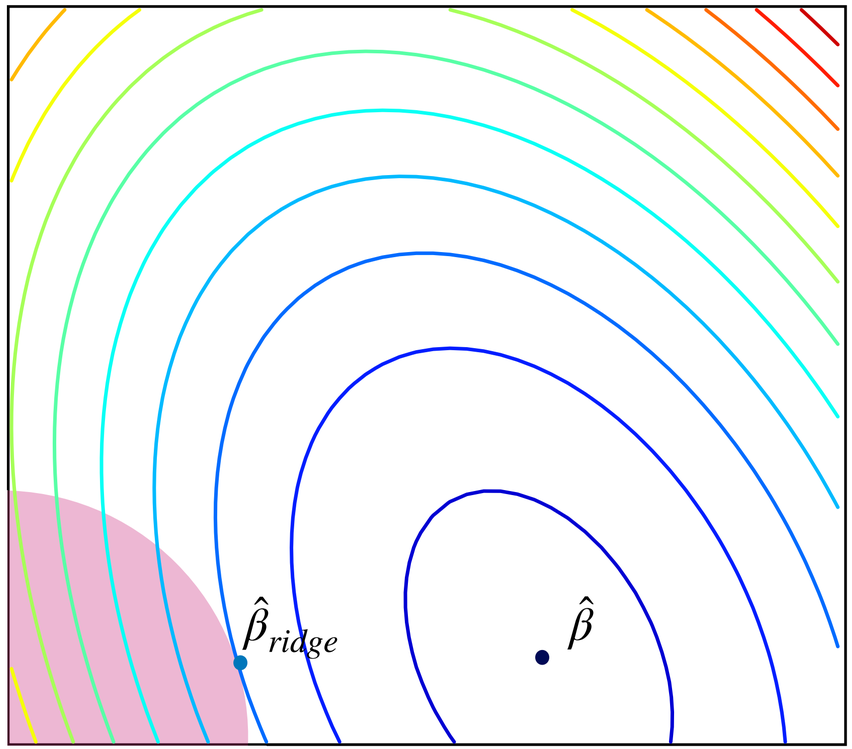
onde $\lambda \geq 0$ depende de R e da amostra $\mathcal{D}$

Este problema não tem uma solução analítica no caso geral!

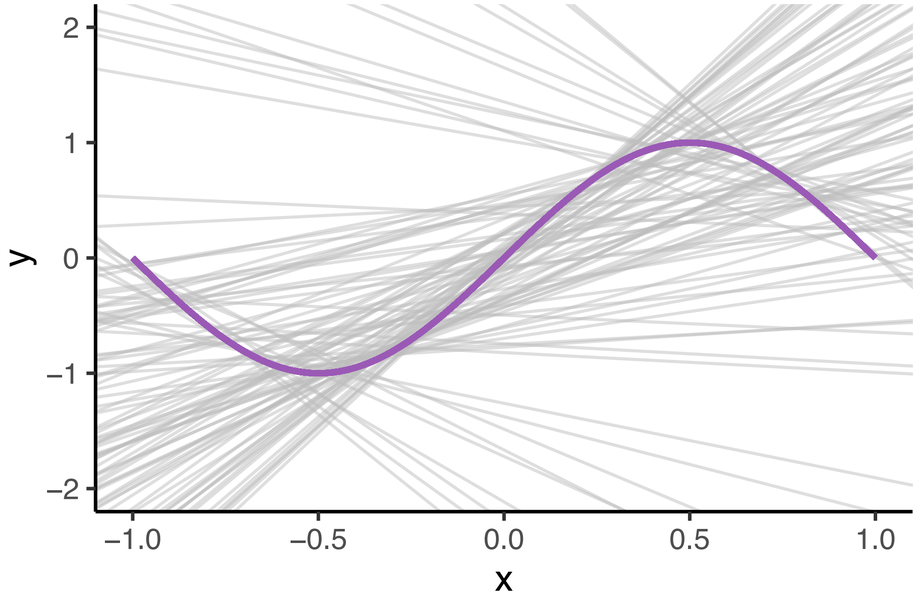
Solução aproximada baseada em métodos de otimização convexa

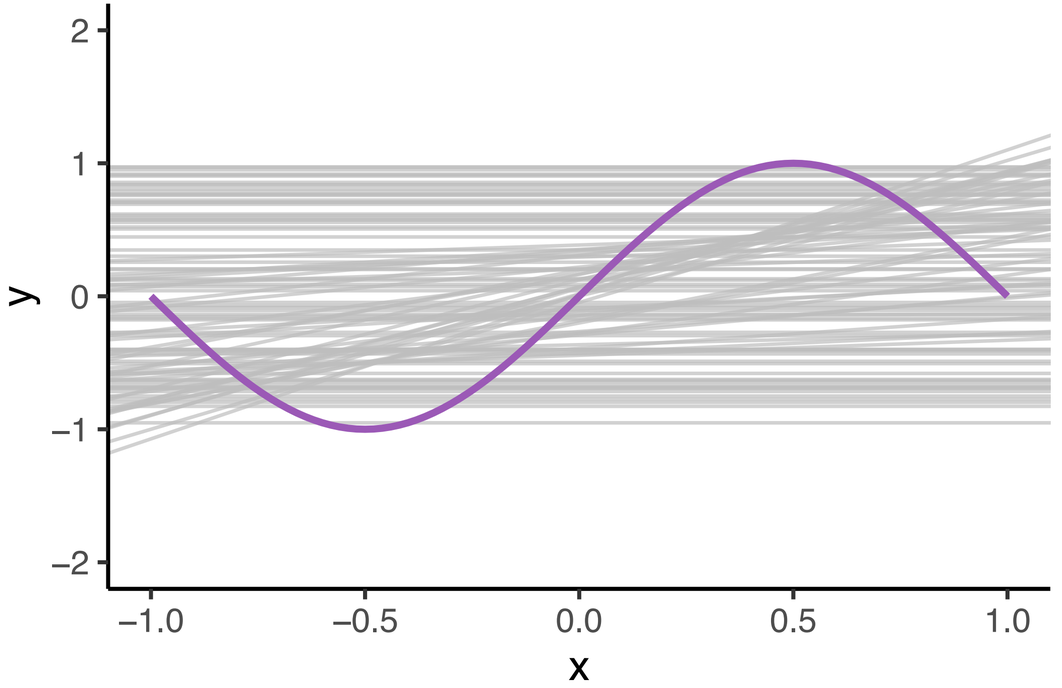
Comparação de RIDGE e LASSO





Exemplo

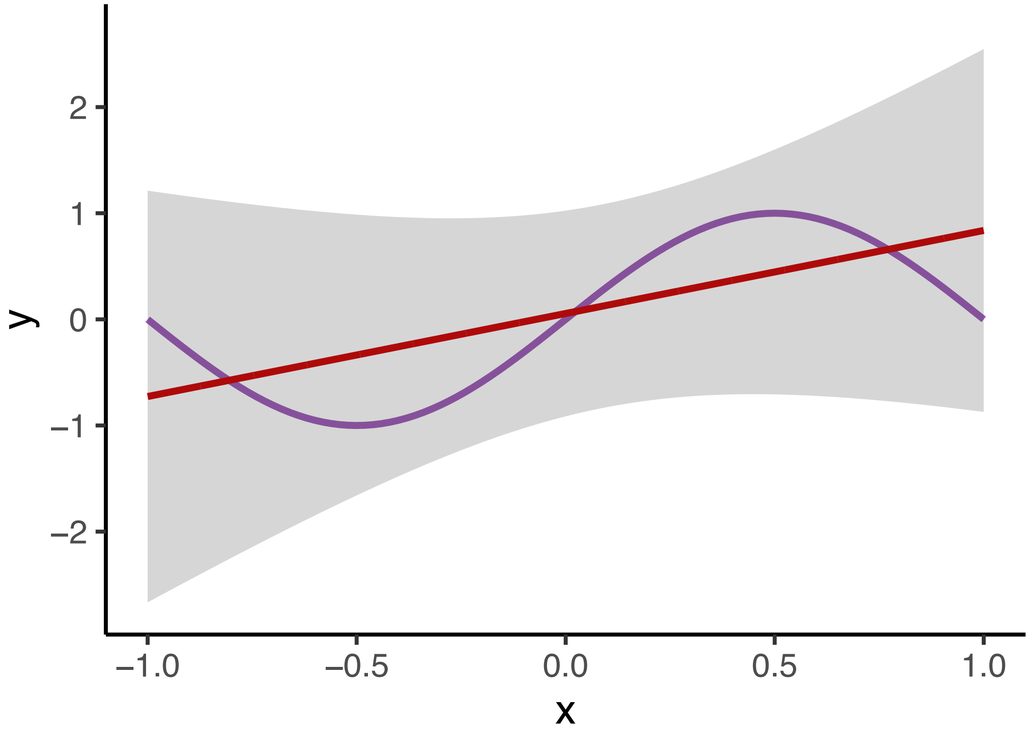




$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x\,, (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

$$\mathcal{G}_{lasso} = \{g(x) = \beta_0 + \beta_1 x,\ \ |\beta_1| \leq R\}$$

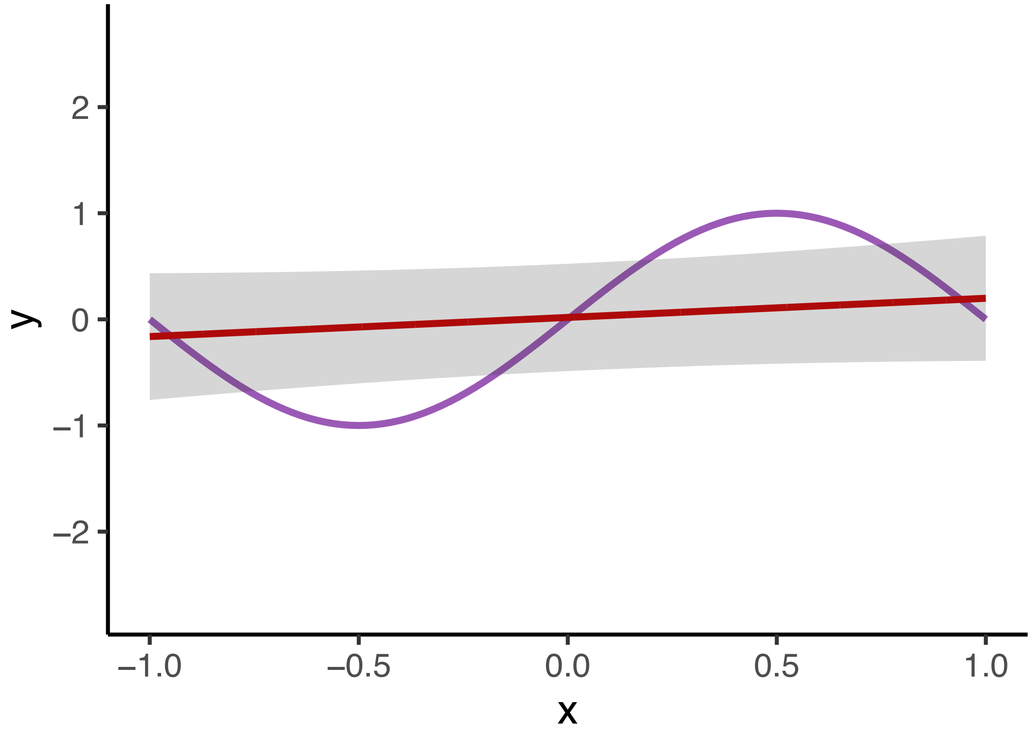
Decomposição do erro em viés e variância



$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x : (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

$$\text{Viés}^2 = 0{,}21$$

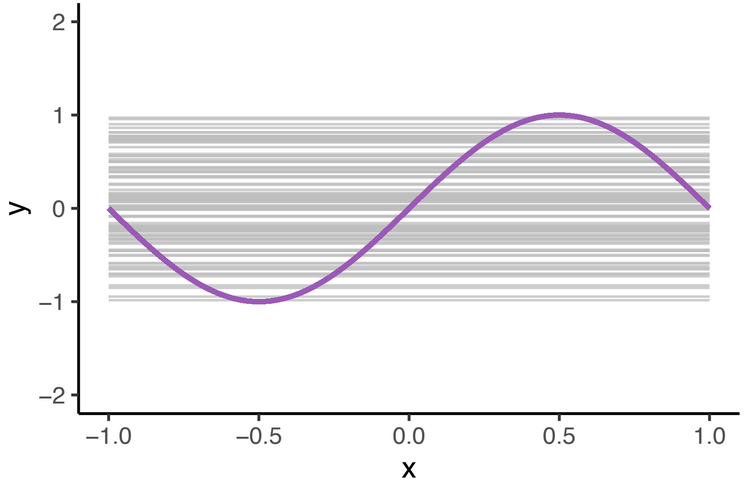
$$\text{Variância} = 0{,}69$$



$$\mathcal{G}_{lasso} = \{g(x) = \beta_0 + \beta_1 x, \ |\beta_1| \leq R\}$$

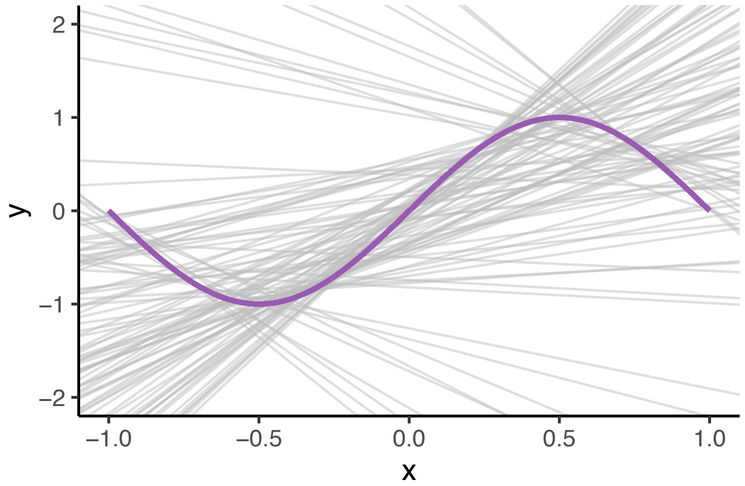
$$\text{Viés}^2 = 0{,}39$$

$$\text{Variância} = 0{,}28$$



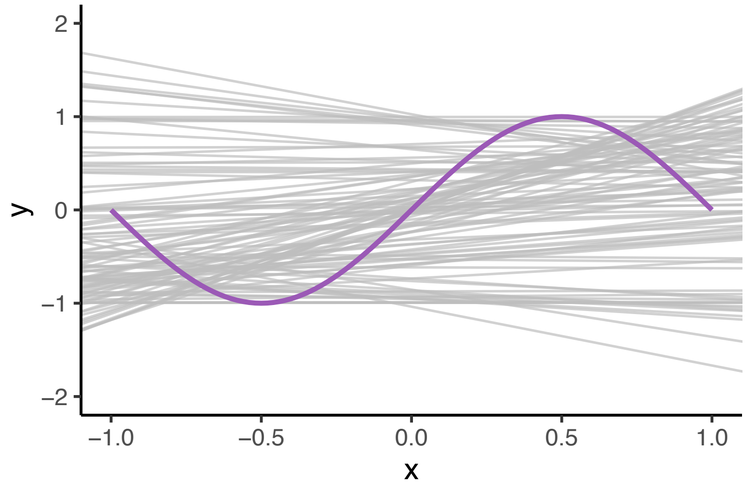
$$\mathcal{G}_1 = \{g(x) = \beta_0\}$$

$$E_F(g^{\mathcal{D}}) = 0{,}75$$



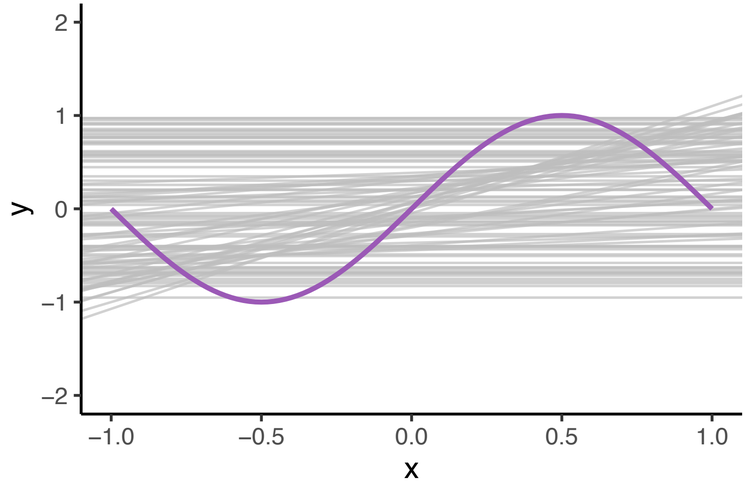
$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x\}$$

$$E_F(g^{\mathcal{D}}) = 1{,}9$$



$$\mathcal{G}_{ridge} = \{g(x) = \beta_0 + \beta_1 x, \ \ |\beta_1|^2 \leq R\}$$

$$E_F(g^{\mathcal{D}}) = 0{,}72$$



$$\mathcal{G}_{lasso} = \{g(x) = \beta_0 + \beta_1 x, \ \ |\beta_1| \leq R\}$$

$$E_F(g^{\mathcal{D}}) = 0{,}67$$

Outras funções de regularização - ELASTIC NET

$$\mathcal{G}_{enet} = \{g(x) = \beta_0 + x^T\beta,\ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \leq R\}$$

Neste modelo, queremos escolher:

$$g \in \mathcal{G}_{enet} \quad \text{que minimize} \quad \widehat{E}_D(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2$$

Isto é equivalente a escolher:

$$\beta \in \mathbb{R}^{p+1} \quad \text{que minimize} \quad \widehat{E}_D^{enet}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2)$$

onde $\lambda \geq 0$ depende de R e da amostra $\mathcal{D}$

Este método é conhecido como “regressão ELASTIC NET”

Mínimos quadrados regularizados - ELASTIC NET

$$\mathcal{G}_{enet} = \{g(x) = \beta_0 + x^T\beta,\ \ \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \leq R\}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\|\beta\|_1 = \sum_{1=1}^{p} |\beta_i|$$

$$\|\beta\|_2^2 = \sum_{1=1}^{p} |\beta_i|^2$$