Aprendizagem estatística em altas dimensões

Florencia Leonardi

Conteúdo

- Comparação de modelos
- Decomposição do erro esperado “fora da amostra”
- Seleção de modelos
- Validação cruzada

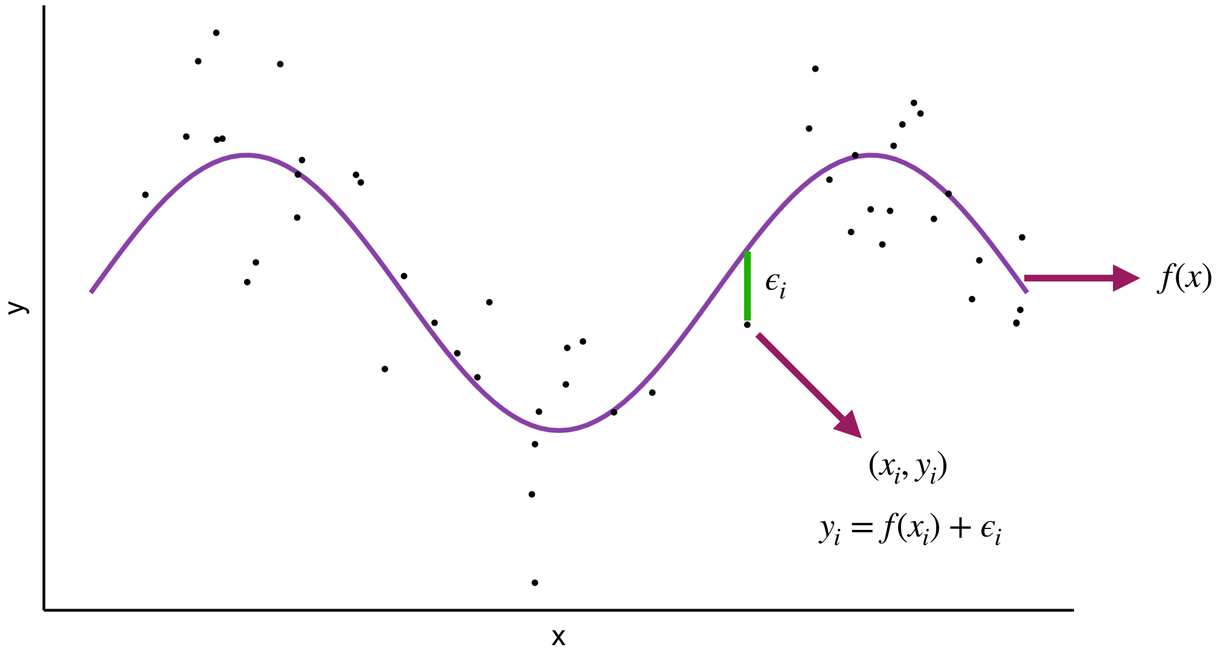
Objetivos da aprendizagem estatística supervisionada

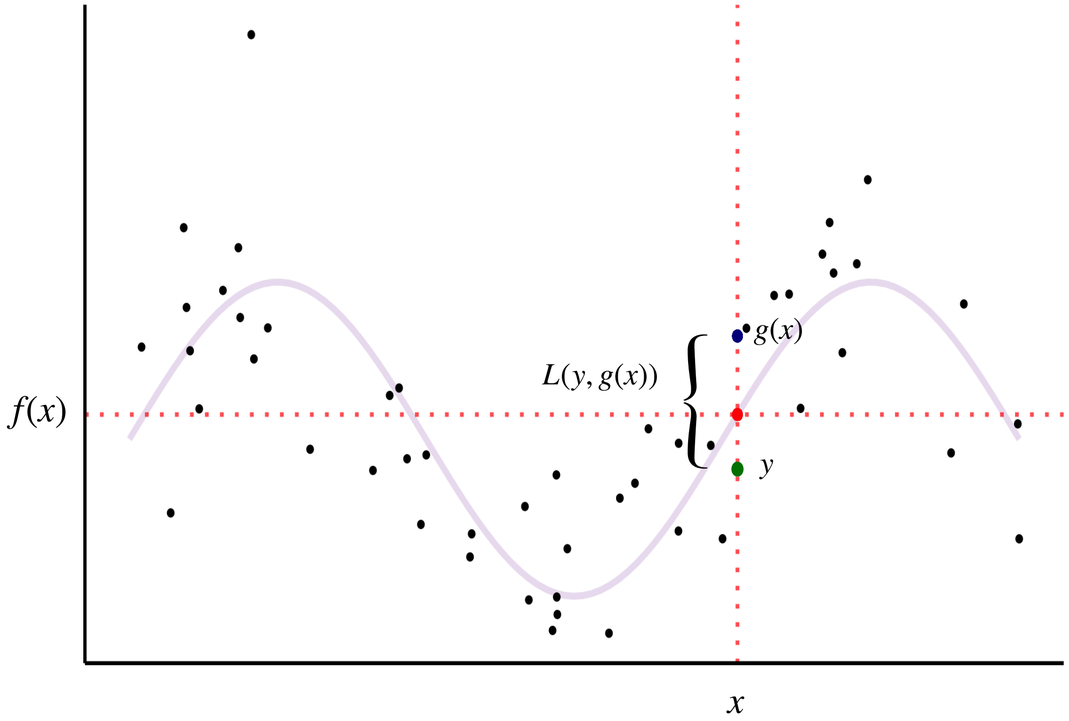
Formalização do problema de aprendizagem estatística supervisionada:

- Uma função objetivo $f\colon \mathcal{X} \to \mathcal{Y}$ desconhecida
- Um conjunto de dados (exemplos) $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Uma função de custo $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$
- Uma família de funções candidatas $\mathcal{G}$ (modelo)

O objetivo da aprendizagem estatística é "aprender" a função objetivo f a partir de um conjunto de dados observado $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

Objetivos da aprendizagem estatística supervisionada





Como escolher g ?

Objetivo: escolher $g \in \mathcal{G}$ que minimize $\mathbb{E}[L(y, g(x))]$

Lembrando:

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ é um estimador de $\mathbb{E}(X)$ se $x_1, \ldots, x_n$ é uma amostra da variável X

Ideia: escolher $g \in \mathcal{G}$ que minimize $\frac{1}{n} \sum_{i=1}^{n} L(y_i, g(x_i))$

Esta ideia funciona bem quando a complexidade da família $\mathcal{G}$ é a adequada para o problema, mas pode ser ruim em vários outros casos !

Tipos de erro

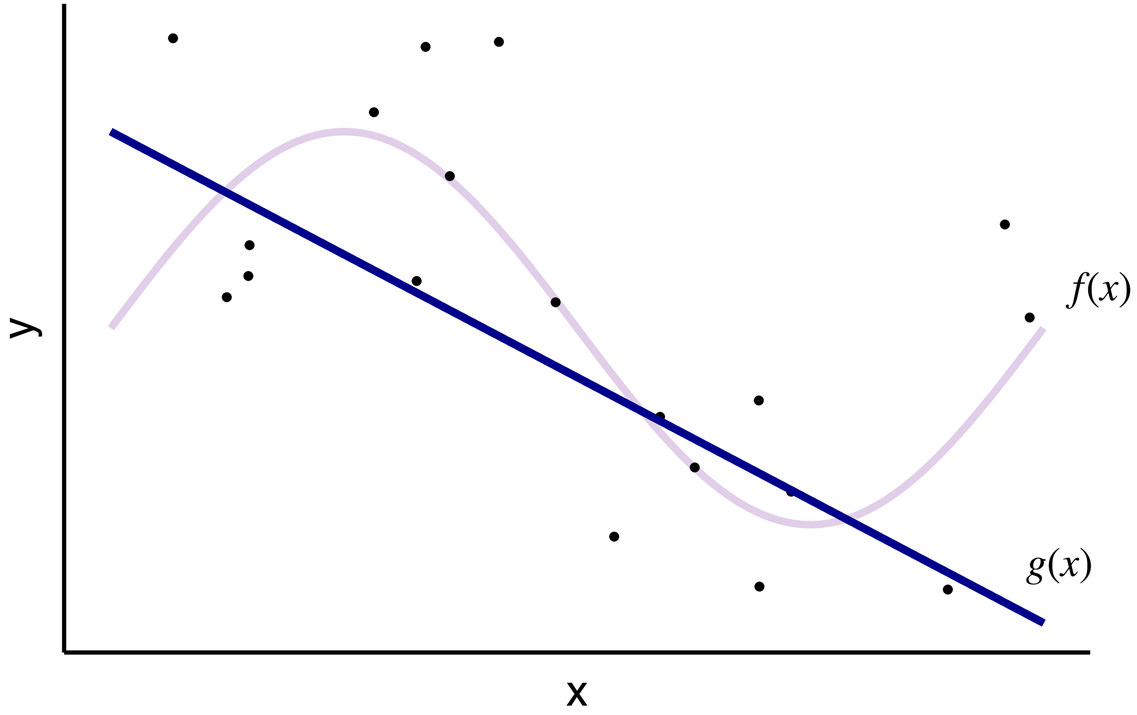
$$\widehat{E}_D(g) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i))$$ → Erro estimado “dentro da amostra"

$$E_D(g) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i))\right]$$ → Erro esperado “dentro da amostra"

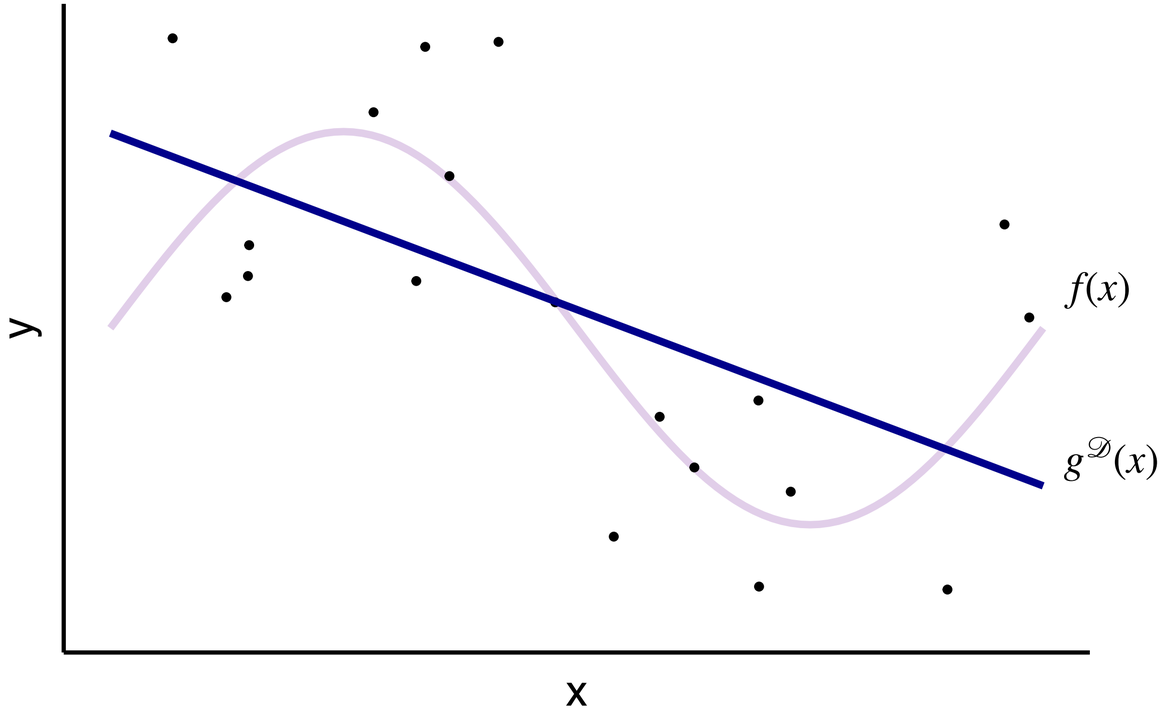
$$E_F(g) = \mathbb{E}(L(y, g(x)))$$ → Erro esperado “fora da amostra"

$$\widehat{E}_F(g)\ ?$$ → Erro estimado “fora da amostra"

A função g que nos interessa depende dos dados!

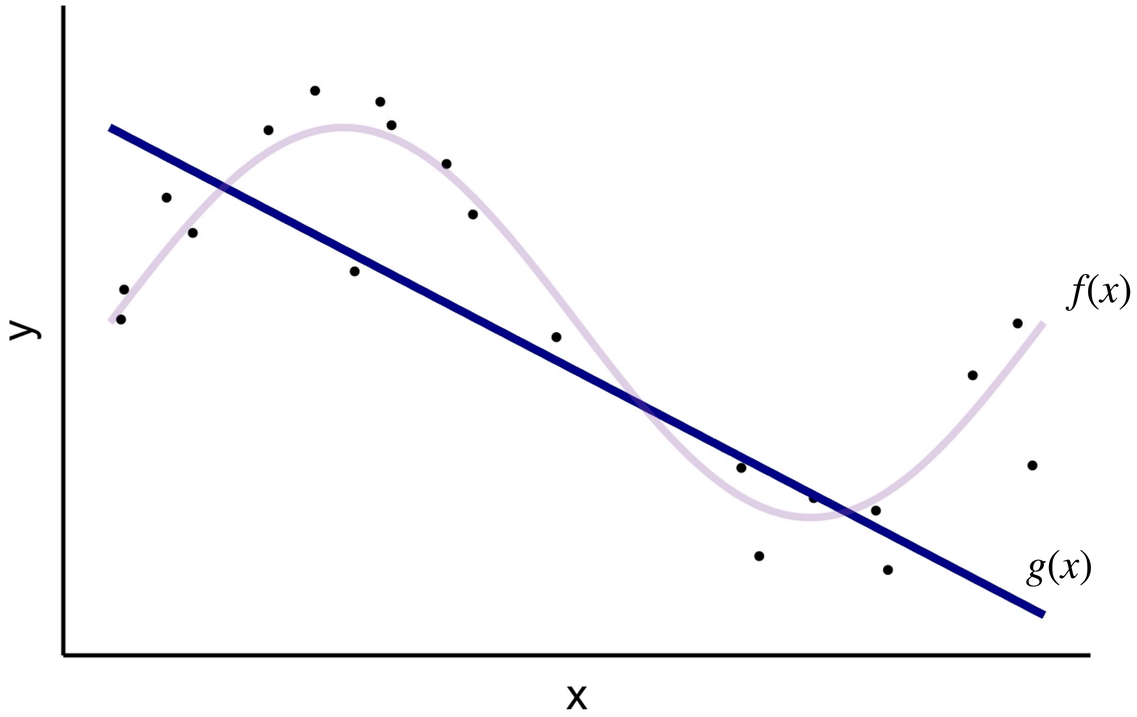


Função g fixa

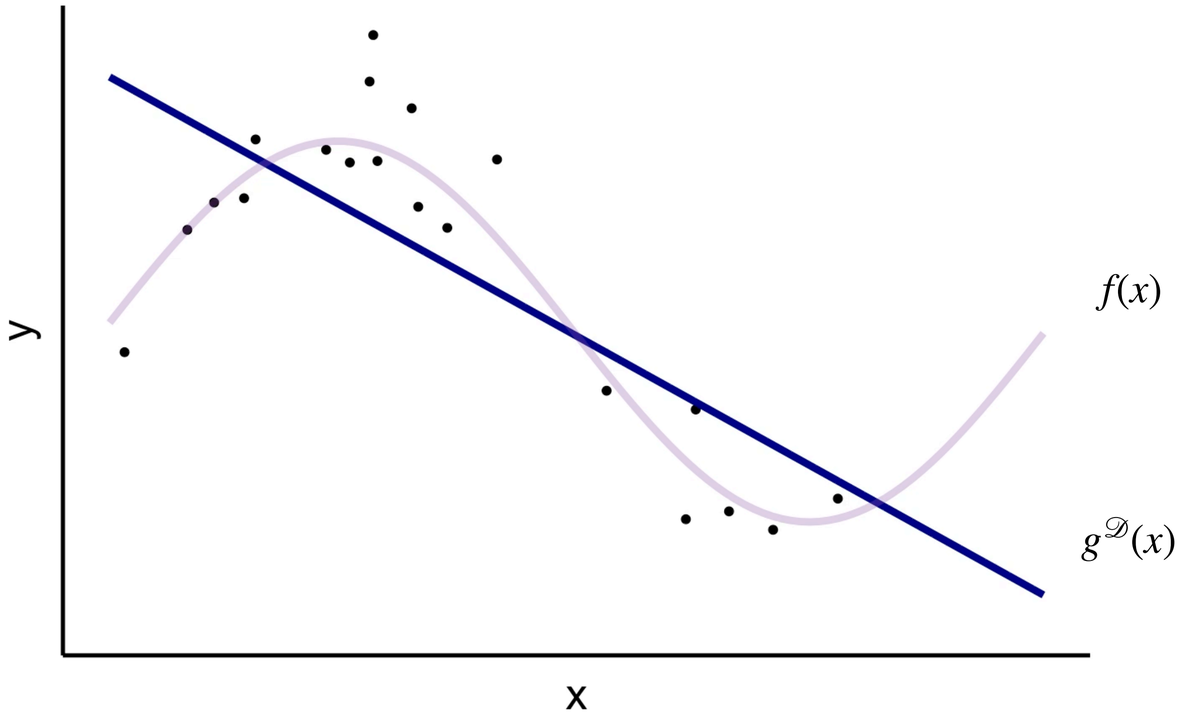


Função g que minimiza $\widehat{E}_D(g)$

A função g que nos interessa depende dos dados!

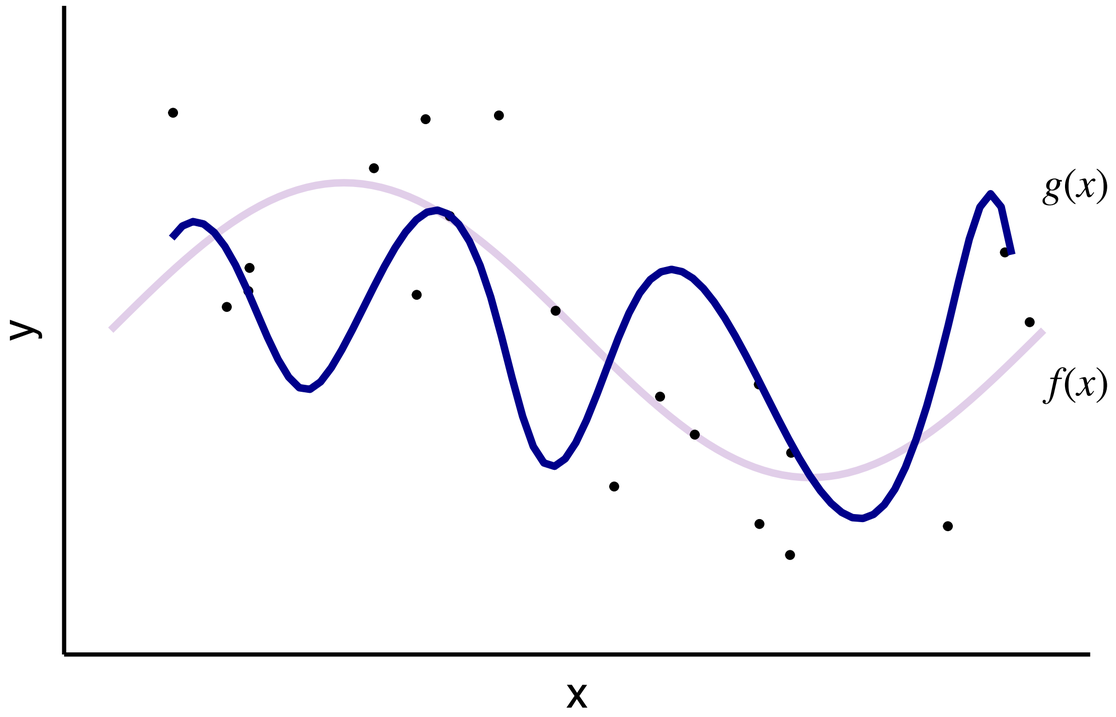


Função g fixa

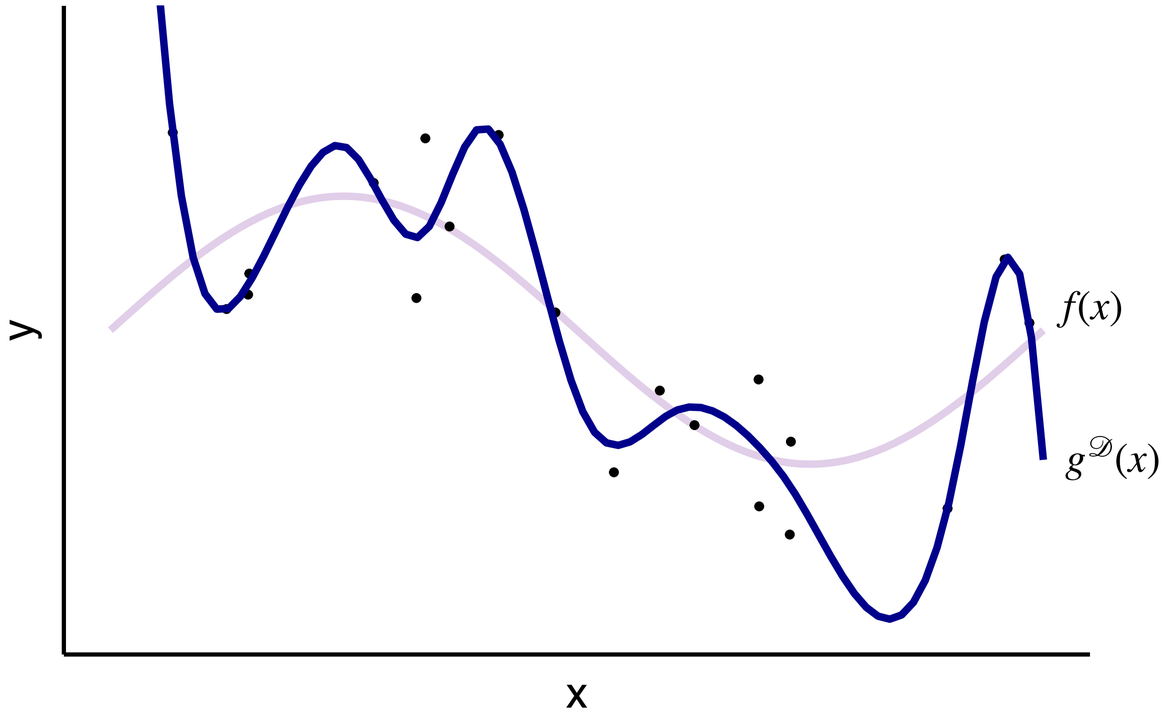


Função g que minimiza $\widehat{E}_D(g)$

A função g que nos interessa depende dos dados!

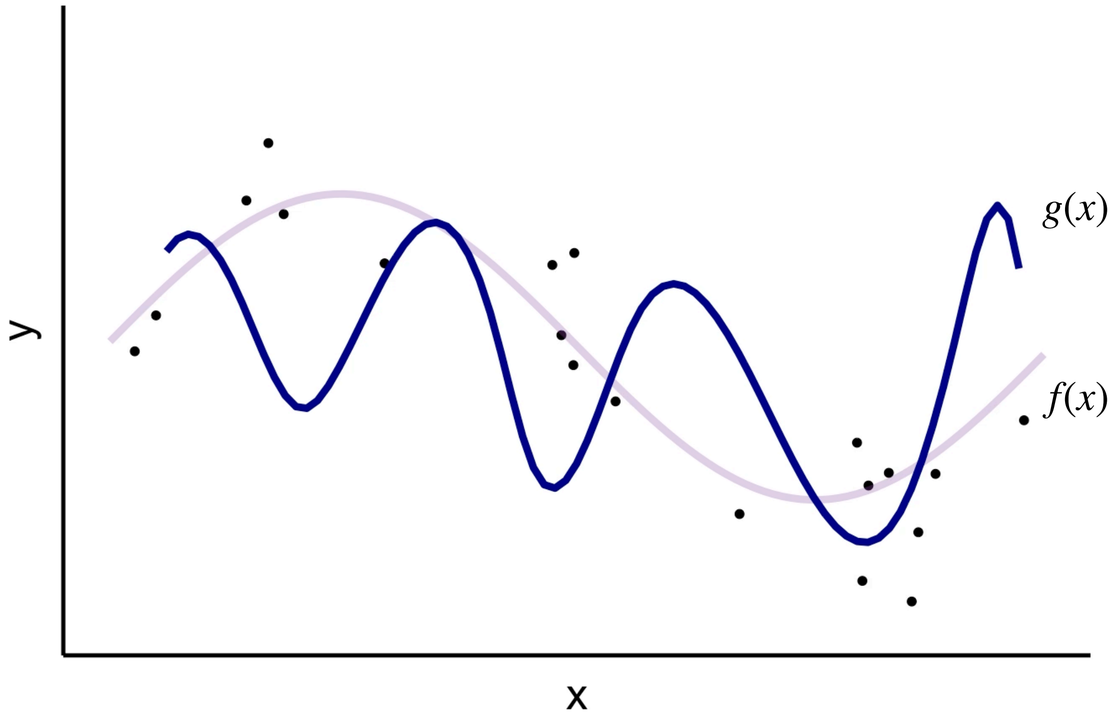


Função g fixa

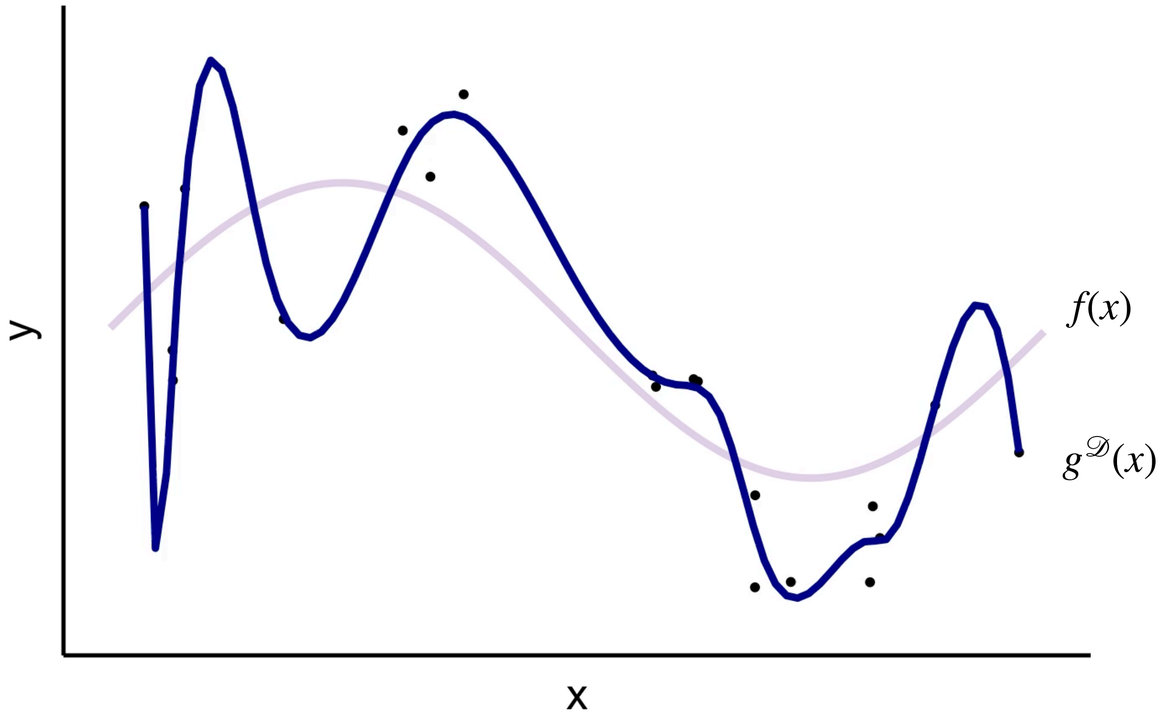


Função g que minimiza $\widehat{E}_D(g)$

A função g que nos interessa depende dos dados!

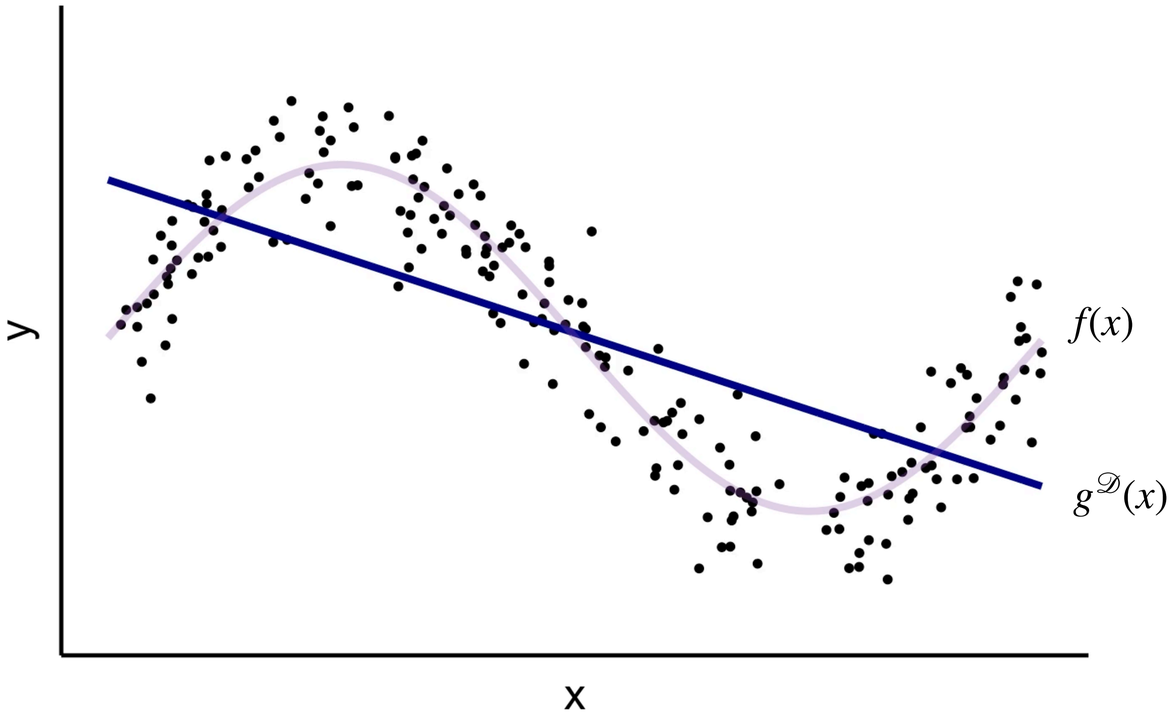


Função g fixa



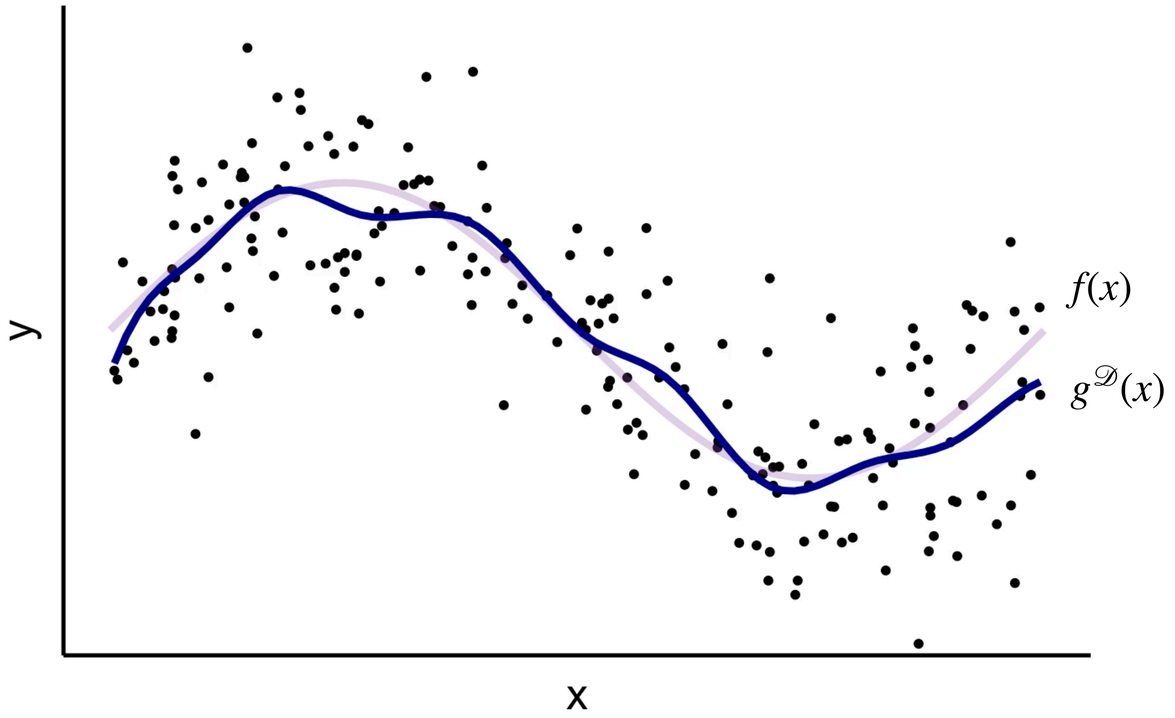
Função g que minimiza $\widehat{E}_D(g)$

E se o tamanho da amostra for maior ... ?



Função g que minimiza $\widehat{E}_D(g)$

$\mathscr{G}$ = funções lineares



Função g que minimiza $\widehat{E}_D(g)$

$\mathscr{G}$ = modelo de splines de grau 12

Decomposição do erro de predição

Para cada $g \in \mathscr{G}$ nós definimos $E_F(g) = \mathbb{E}[L(y, g(x))]$

Mas nós queremos avaliar o erro esperado fora da amostra da função $g^{\mathscr{D}}$, escolhida minimizando $\widehat{E}_D(g)$!

Ou seja nós gostaríamos de saber quanto vale $E_F(g^{\mathscr{D}}) = \mathbb{E}(L(y, g^{\mathscr{D}}(x)))$

A esperança é calculada em relação a uma distribuição de probabilidade.... qual neste caso?

- A distribuição dos dados $\mathscr{D}$ que são i.i.d com distribuição $p(x, y)$
- A distribuição da observação de teste (x, y) que também tem distribuição $p(x, y)$ e é independente de $\mathscr{D}$

Decomposição do erro de predição

Consideremos a função de custo quadrática $L(y, \hat{y}) = (y - \hat{y})^2$ no caso de regressão

Dada $g \in \mathcal{G}$ podemos escrever:

$$E_F(g) = \mathbb{E}[(f(x) + \epsilon - g(x))^2] = \mathbb{E}[(f(x) - g(x))^2 + 2(f(x) - g(x))\epsilon + \epsilon^2]$$

$$= \mathbb{E}[(f(x) - g(x))^2] + \mathbb{E}[2(f(x) - g(x))\epsilon] + \mathbb{E}[\epsilon^2]$$

Como assumimos que as variáveis X e ϵ são independentes e ainda $\mathbb{E}[\epsilon] = 0$, podemos verificar que

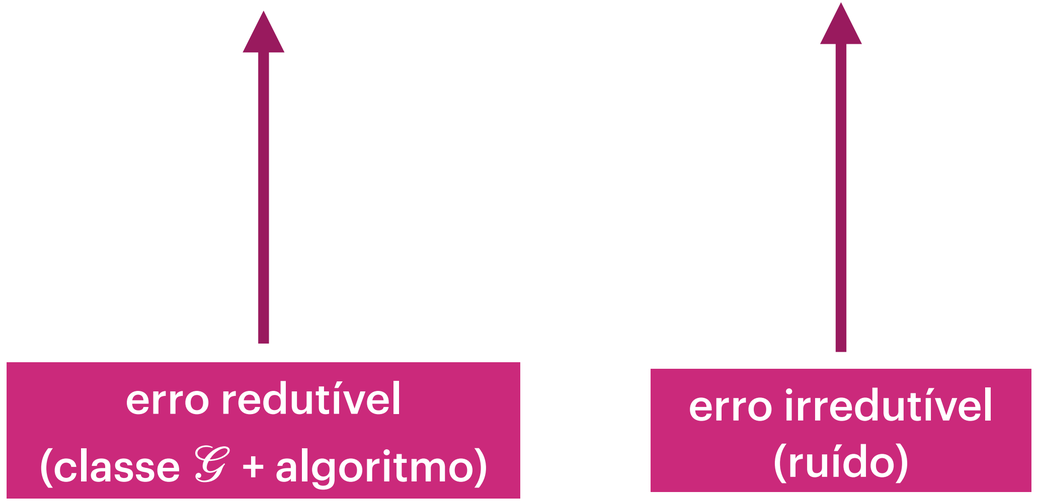
$$\mathbb{E}(2(f(x) - g(x))\epsilon) = 2\mathbb{E}[f(x) - g(x)]\mathbb{E}[\epsilon] = 0$$

Na formulação do modelo ainda assumimos que $\mathbb{E}[\epsilon^2] = \sigma^2$, logo obtemos que

$$E_F(g) = \mathbb{E}[(f(x) - g(x))^2] + \sigma^2$$

Decomposição do erro de predição

$$E_F(g) = \mathbb{E}[(f(x) - g(x))^2] + \sigma^2$$



Decomposição do erro de predição

Consideremos agora o erro redutível

$$\mathbb{E}[(f(x) - g^{\mathscr{D}}(x))^2] \;=\; \mathbb{E}_{(x,y)}\mathbb{E}_{\mathscr{D}}[(f(x) - g^{\mathscr{D}}(x))^2]$$

Para um x fixo, podemos pensar em $g^{\mathscr{D}}(x)$ como um estimador de $f(x)$

Aqui, $g^{\mathscr{D}}$ é uma função obtida por um método qualquer, não necessariamente a que minimiza $\widehat{E}_D(g)$

Lembrando:

$$\text{EQM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] \;=\; \text{Viés}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

Logo:

$$\mathbb{E}_{\mathscr{D}}[(f(x) - g^{\mathscr{D}}(x))^2] \;=\; \text{Viés}(g^{\mathscr{D}}(x))^2 + \text{Var}(g^{\mathscr{D}}(x))$$

Decomposição do erro de predição

$$\mathbb{E}_{\mathcal{D}}[(f(x) - g^{\mathcal{D}}(x))^2] = \text{Viés}(g^{\mathcal{D}}(x))^2 + \text{Var}(g^{\mathcal{D}}(x))$$

$$\text{Viés}(g^{\mathcal{D}}(x))^2 = [\mathbb{E}_{\mathcal{D}}(g^{\mathcal{D}}(x)) - f(x)]^2$$

$$\text{Var}(g^{\mathcal{D}}(x)) = \mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \underbrace{\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]}_{\bar{g}(x)})^2]$$

→ Muitas vezes $\bar{g}(x)$ é a melhor aproximação de f dentro da classe $\mathcal{G}$

$$g^* = \inf_{g \in \mathcal{G}} \mathbb{E}_x[(f(x) - g(x))^2]$$

Exemplo

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

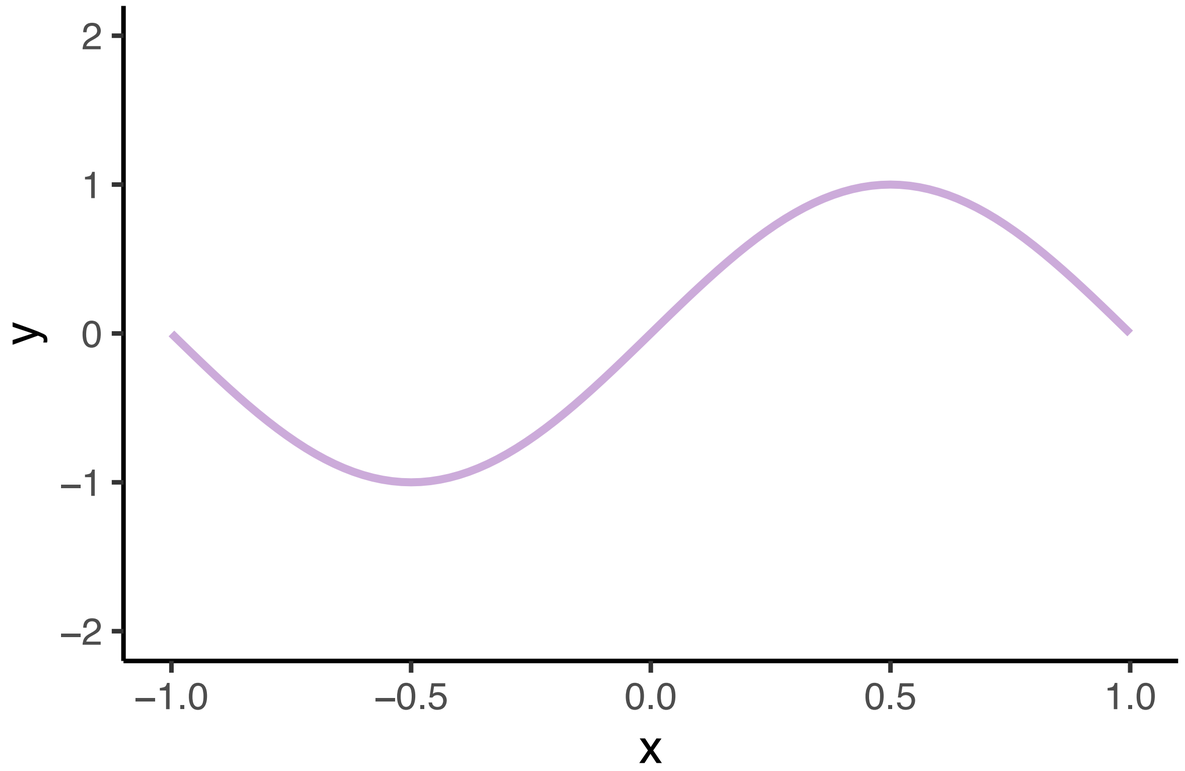
$f(x) = \sin(\pi x)$

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

Duas classes de modelos

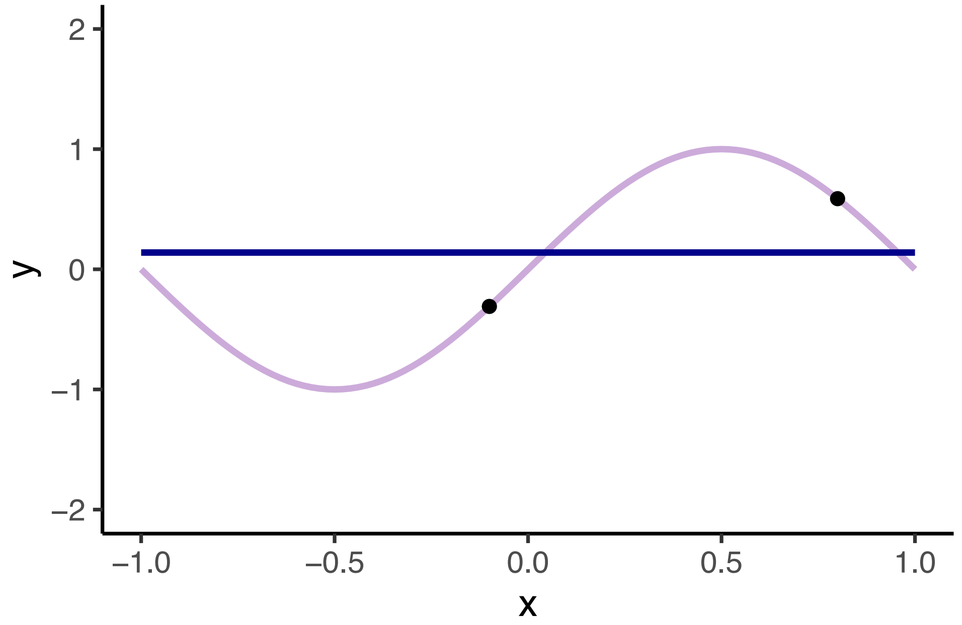
$$\mathcal{G}_1 = \{g(x) = \beta_0 \colon \beta_0 \in \mathbb{R}\}$$

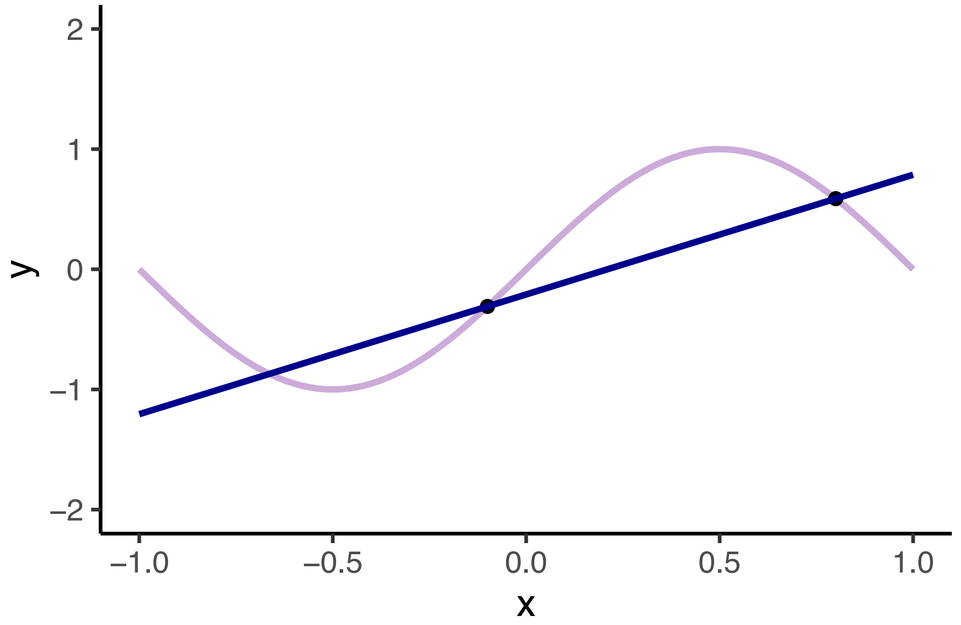
$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x \colon (\beta_0, \beta_1) \in \mathbb{R}^2\}$$



Exemplo

Suponhamos que nosso conjunto de dados $\mathscr{D}$ só tem dois pontos:



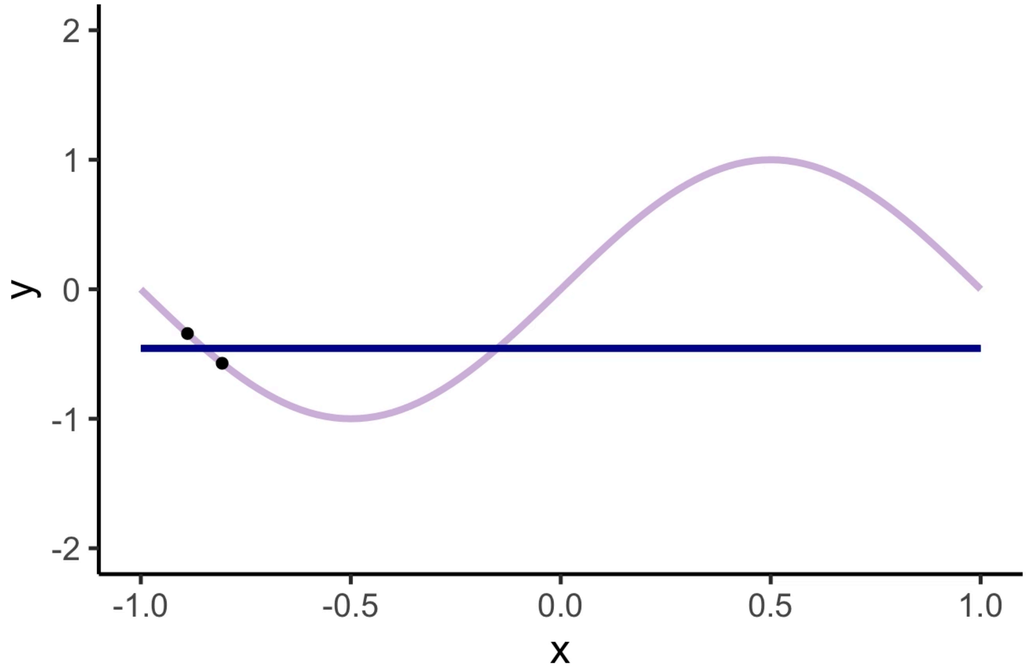


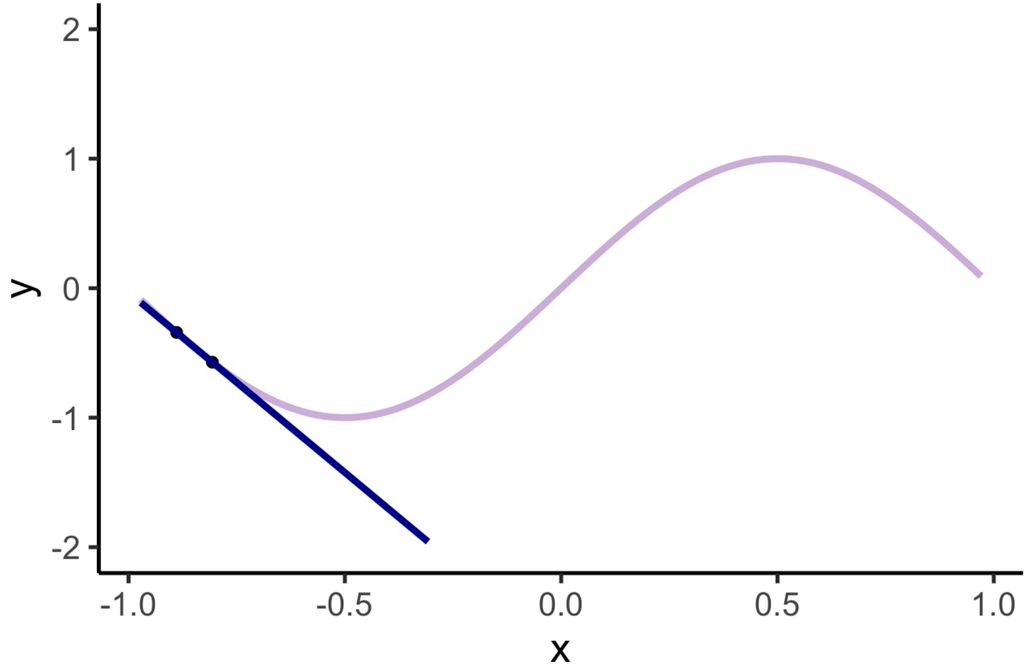
$$\mathscr{G}_1 = \{g(x) = \beta_0 \colon \beta_0 \in \mathbb{R}\}$$

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

Exemplo

Suponhamos que nosso conjunto de dados $\mathscr{D}$ só tem dois pontos:

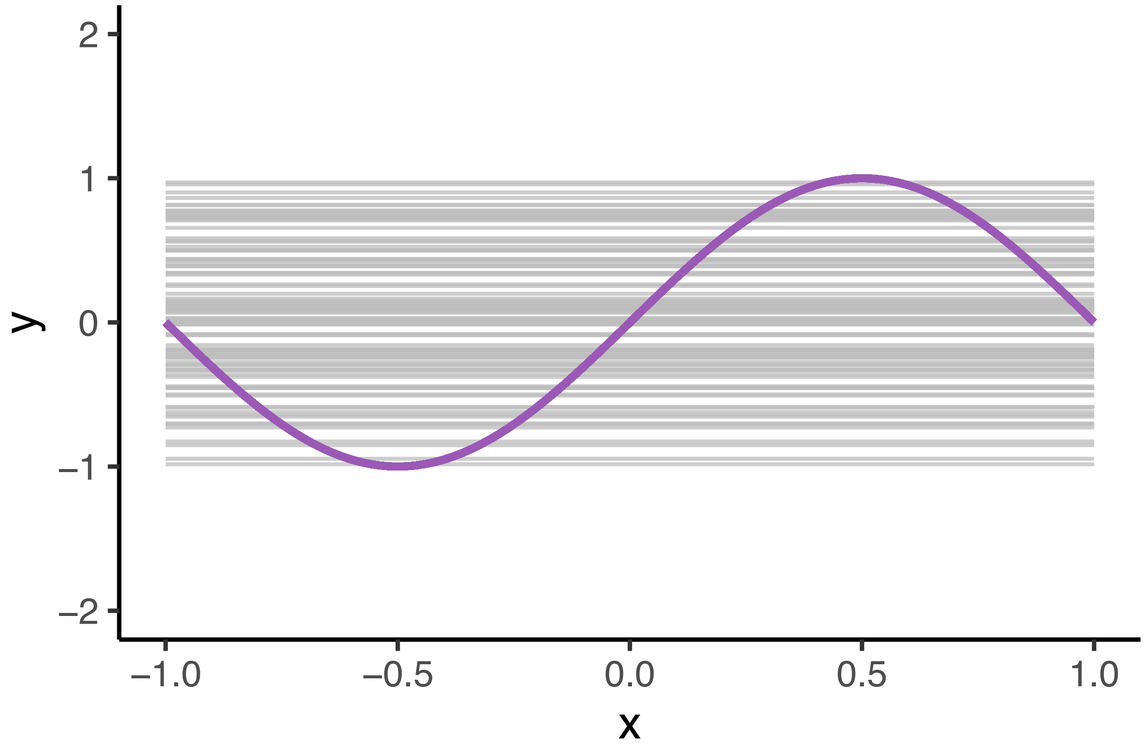


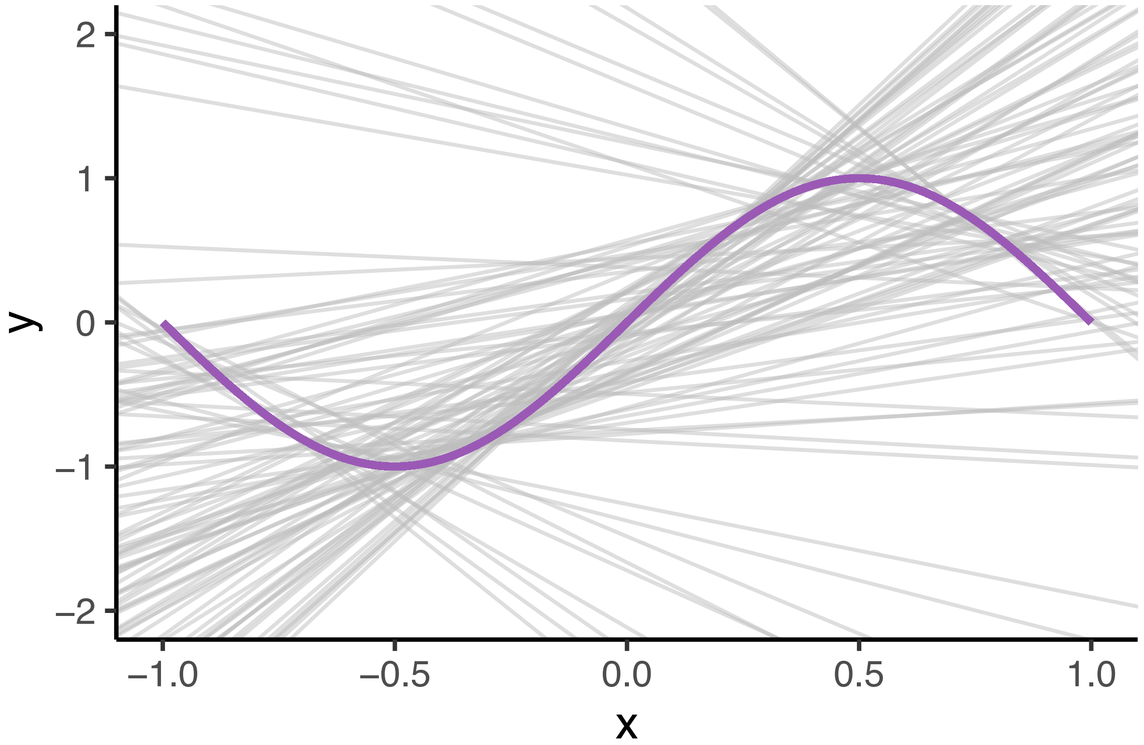


$$\mathscr{G}_1 = \{g(x) = \beta_0 \colon \beta_0 \in \mathbb{R}\}$$

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

Exemplo

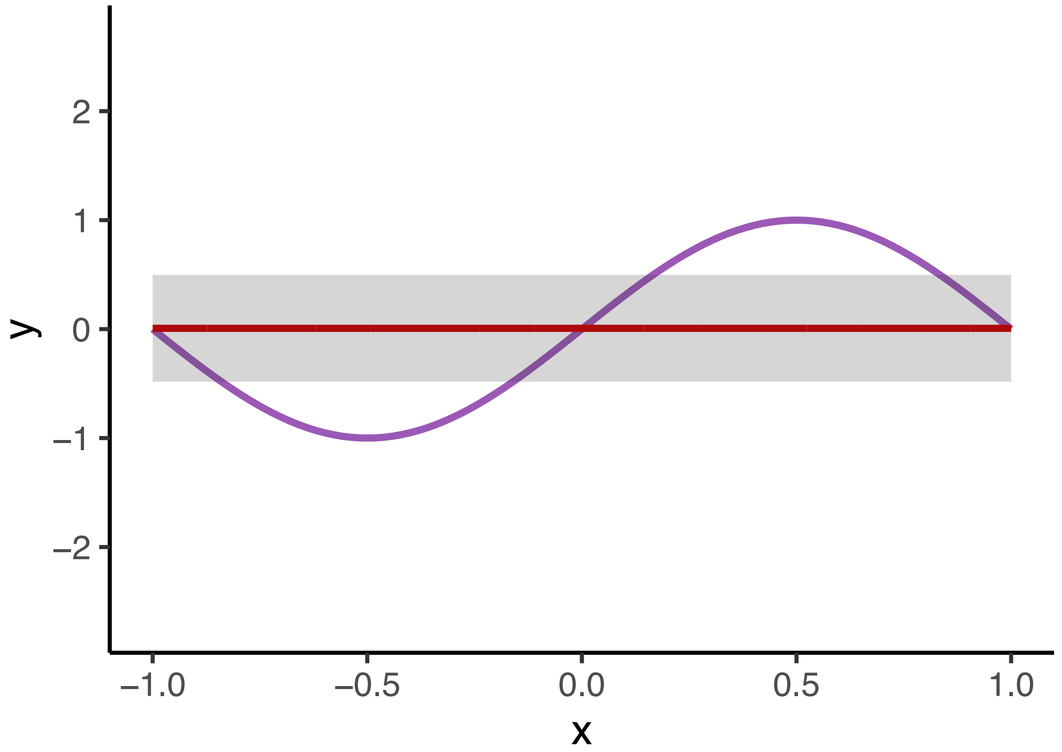




$$\mathcal{G}_1 = \{g(x) = \beta_0 \colon \beta_0 \in \mathbb{R}\}$$

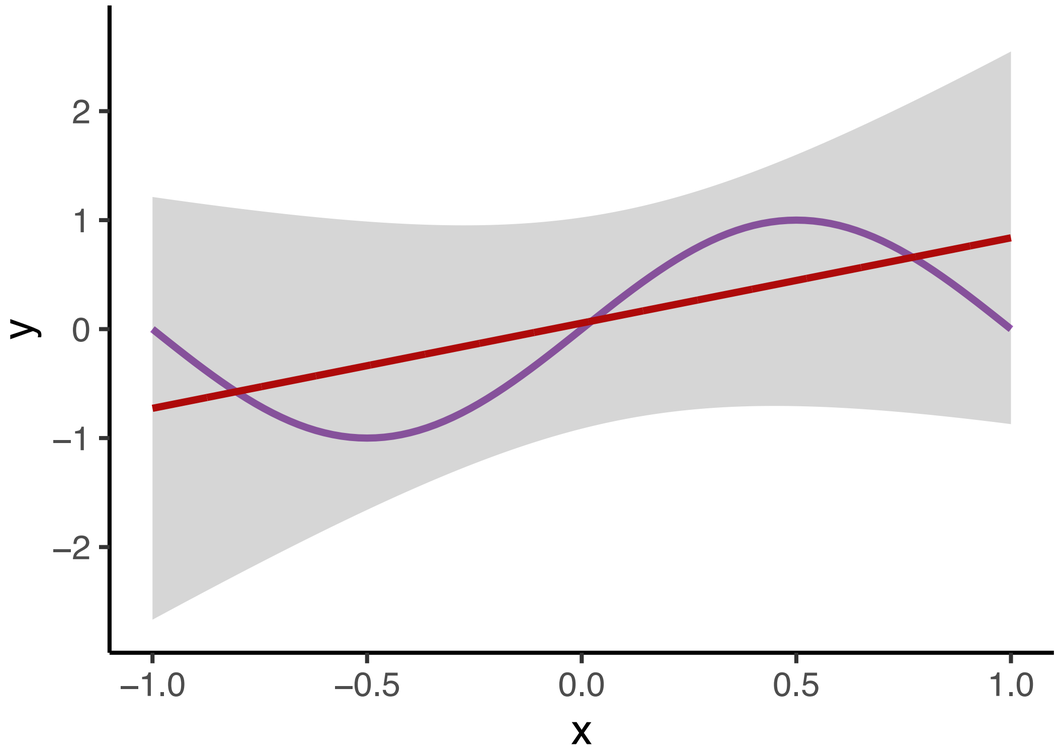
$$\mathcal{G}_2 = \{g(x) = \beta_0 + \beta_1 x \colon (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

Exemplo



$$\mathscr{G}_1 = \{g(x) = \beta_0 : \beta_0 \in \mathbb{R}\}$$

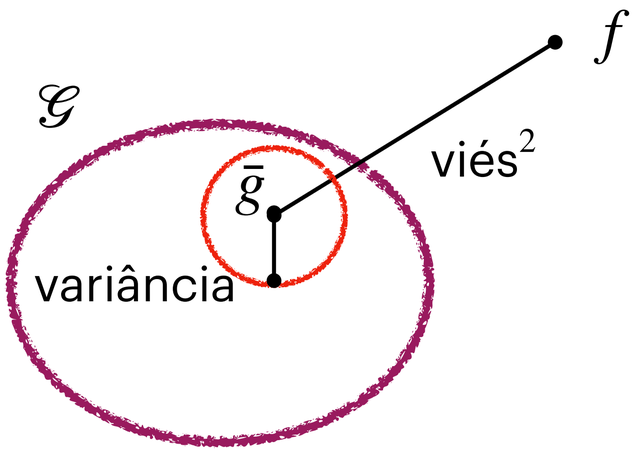
$\text{Viés}^2 = 0{,}50$ $\quad$ $\text{Variância} = 0{,}25$



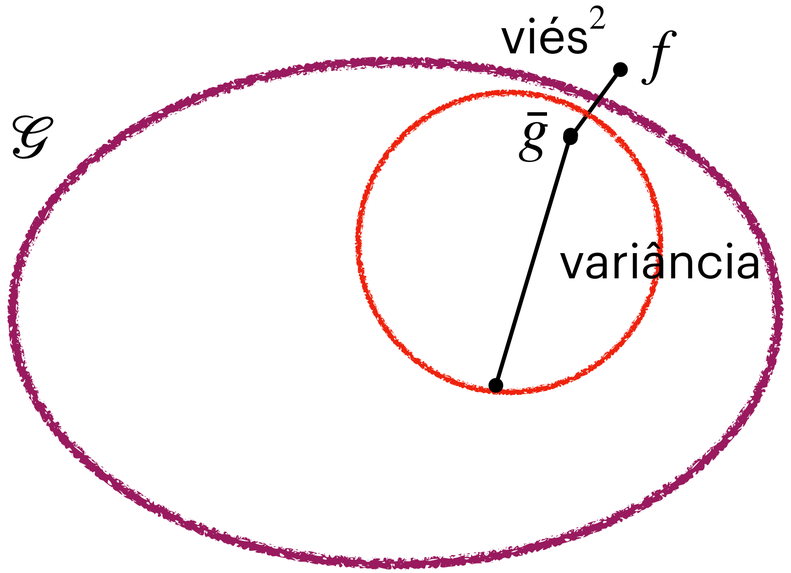
$$\mathscr{G}_2 = \{g(x) = \beta_0 + \beta_1 x : (\beta_0, \beta_1) \in \mathbb{R}^2\}$$

$\text{Viés}^2 = 0{,}21$ $\quad$ $\text{Variância} = 1{,}69$

O custo benefício

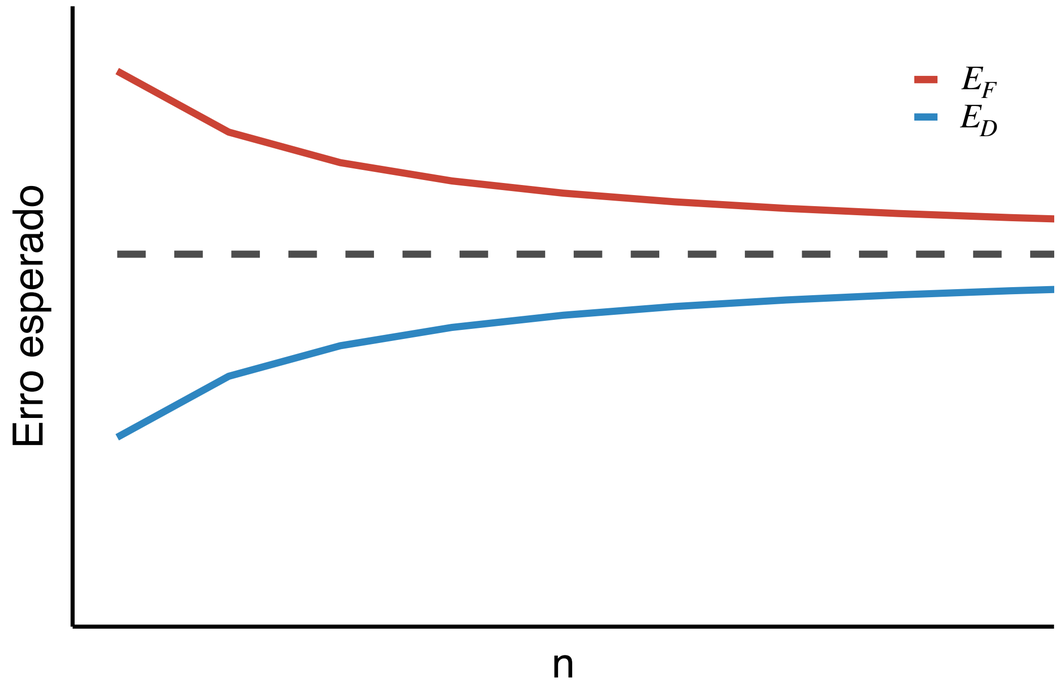


Modelo “simples”

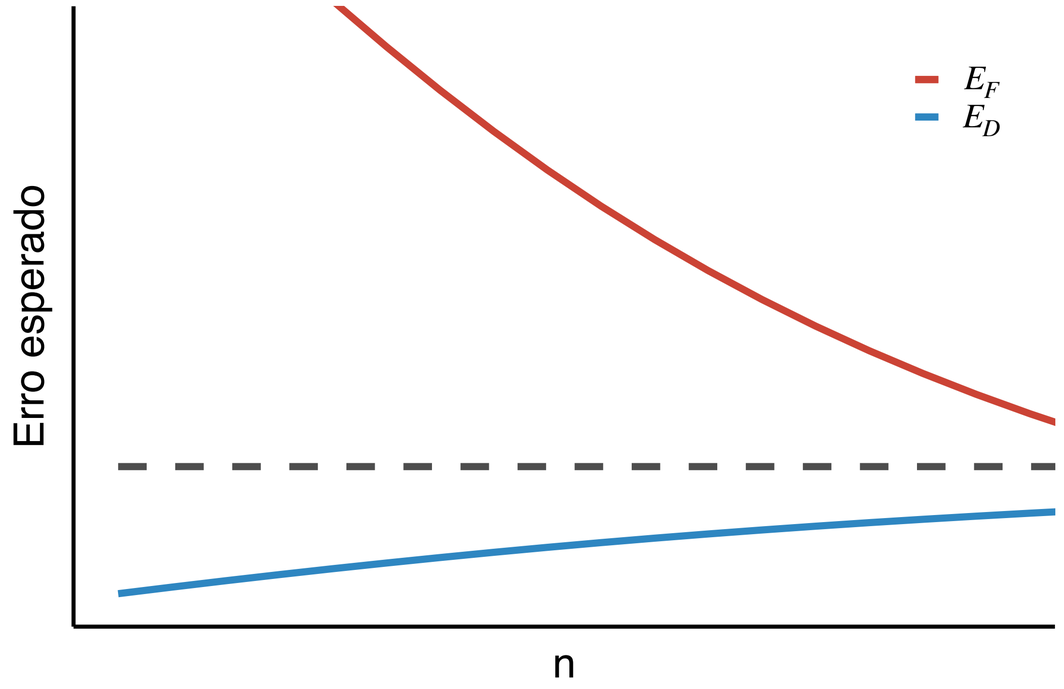


Modelo “complexo”

Curvas de erro

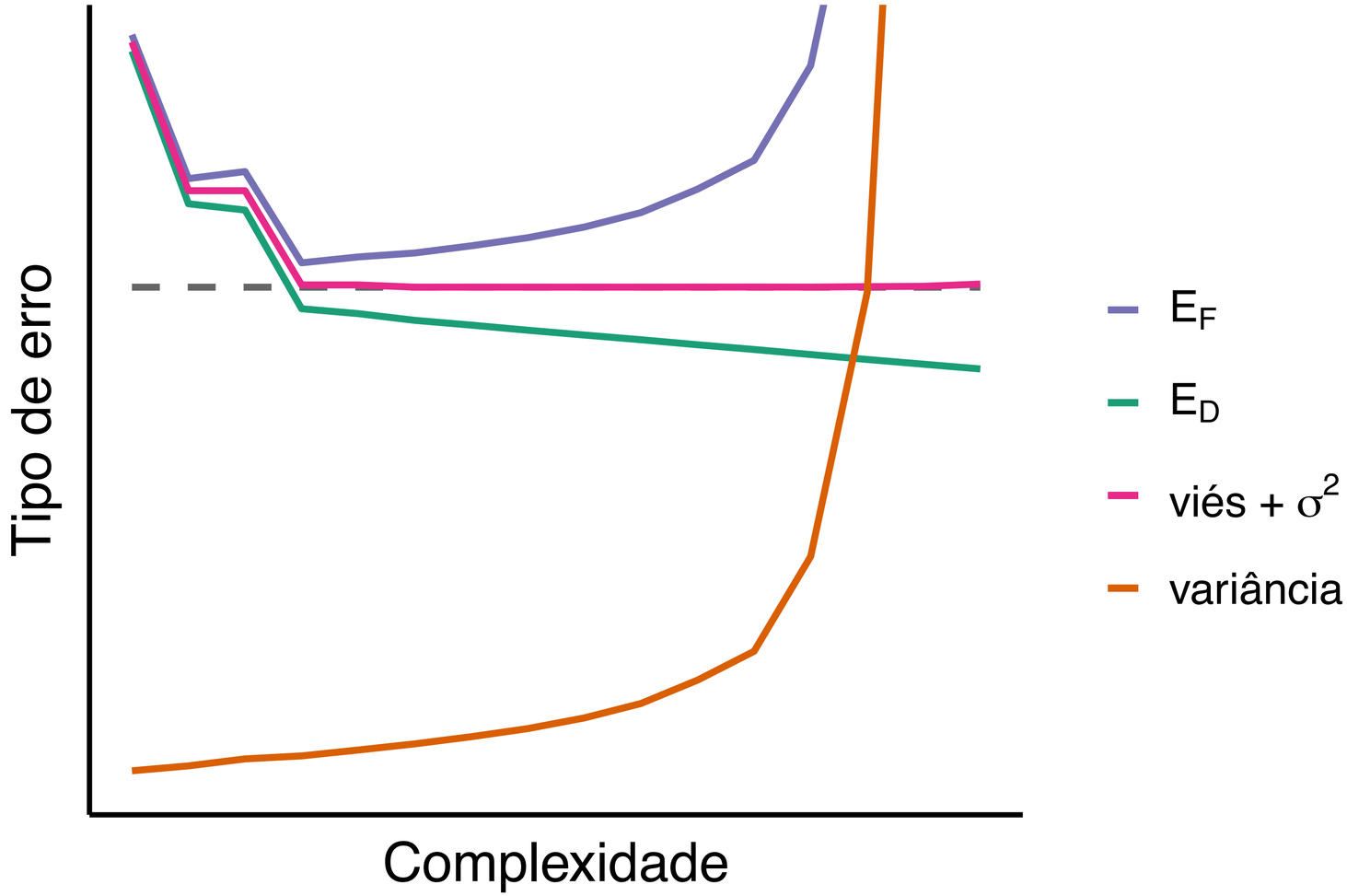


Modelo “simples”



Modelo “complexo”

Curvas de erro



O custo benefício

“A complexidade do modelo deve estar de acordo com os dados disponíveis, e não com a complexidade da função objetivo”

Yaser Abu Mostafa - Learning from data - MIT Online Course

Estimação do erro “fora da amostra”

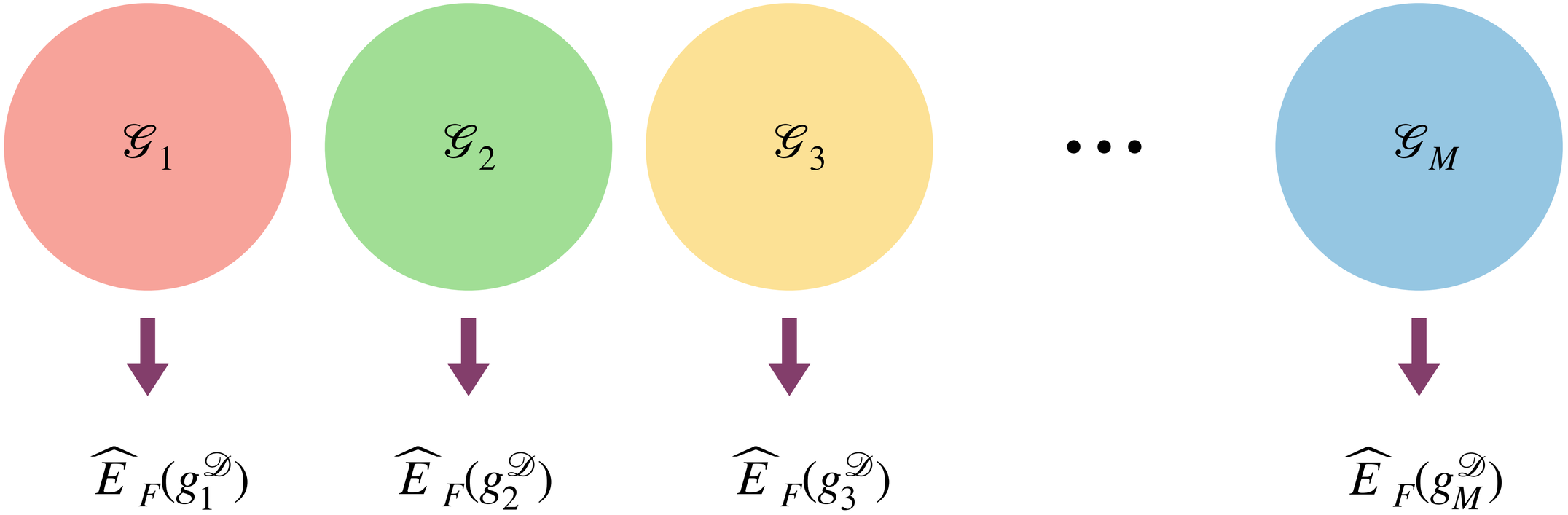
Se nós tivermos acesso a uma amostra independente de $\mathscr{D}$, que denotamos por $\mathscr{D}_{Te} = \{(x_1, y_1), \ldots, (x_K, y_K)\}$, podemos estimar $E_F(g) = \mathbb{E}(L(y, g))$ pela média empírica $\widehat{E}_F(g) = \frac{1}{K}\sum_{k=1}^{K} L(y_i, g(x_i))$

Neste caso, mesmo para $g = g^{\mathscr{D}}$ temos que

$$\left.\begin{aligned} &\mathbb{E}(\widehat{E}_F(g)) = E_F(g) \\ &\text{Var}(\widehat{E}_F(g)) = \frac{\text{Var}[L(y, g(x))]}{K} = \frac{C}{K} \end{aligned}\right\} \rightarrow E_F(g) = \widehat{E}_F(g) \pm O_p\left(\frac{1}{\sqrt{K}}\right)$$

Estimação do erro “fora da amostra”

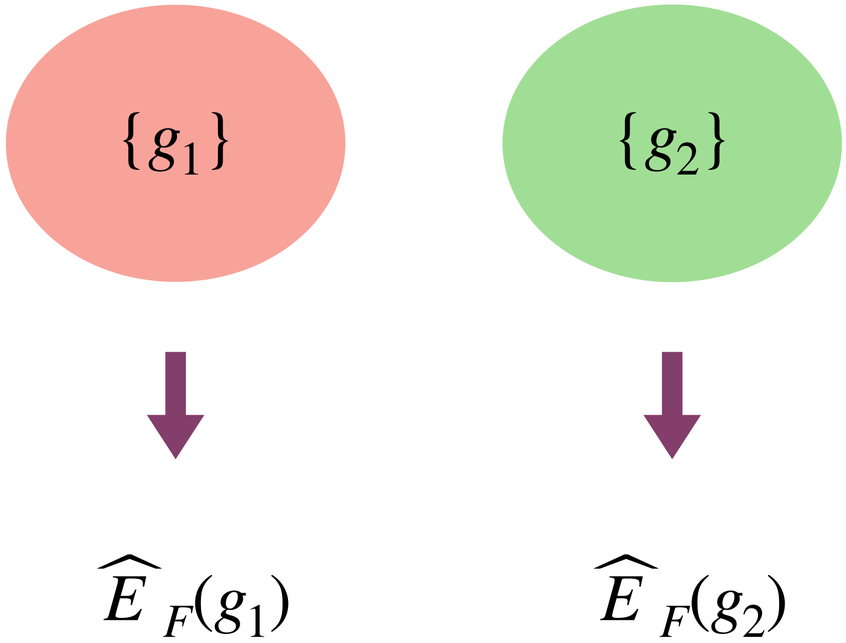
Seleção de modelos



Escolhemos $g_i^{\mathcal{D}}$ tal que $\widehat{E}_F(g_i^{\mathcal{D}})$ é mínimo?

Seleção de modelos

Escolhemos $g_i^{\mathscr{D}}$ tal que $\widehat{E}_F(g_i^{\mathscr{D}})$ é mínimo?



$$\widehat{E}_F(g_1),\ \widehat{E}_F(g_2) \sim \text{Uniforme}(0,1)$$

$$E_F(g_1) = E_F(g_2) = 0.5$$

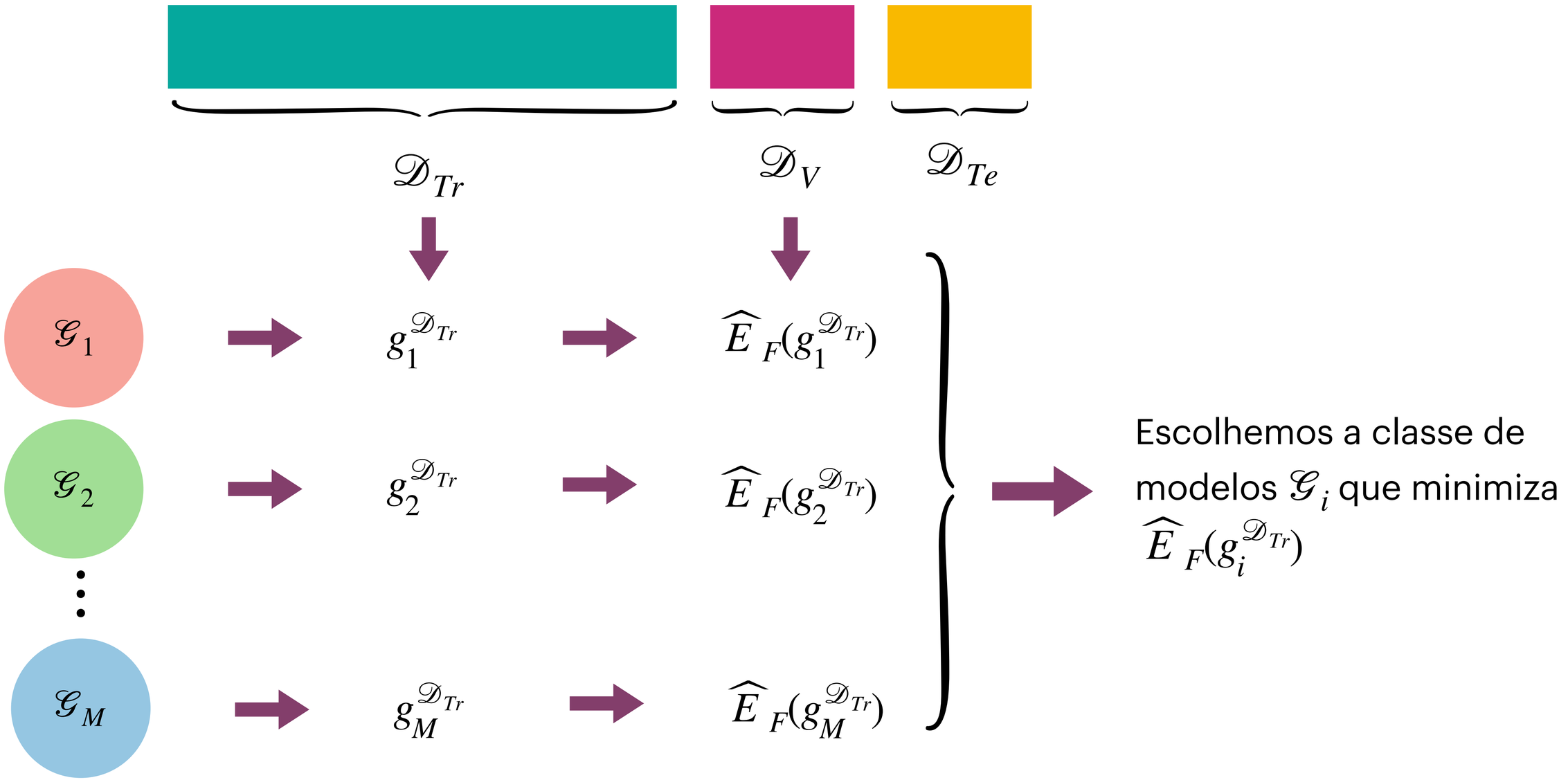
$$g^{\mathscr{D}} = \arg\min(\widehat{E}_F(g_1),\ \widehat{E}_F(g_2))$$

$$\mathbb{E}[\widehat{E}_F(g^{\mathscr{D}})] < 0.5\ !$$

Seleção de modelos

Quando um conjunto de dados de teste é utilizado para guiar a escolha do modelo, então ele passa a ser um conjunto de "validação"

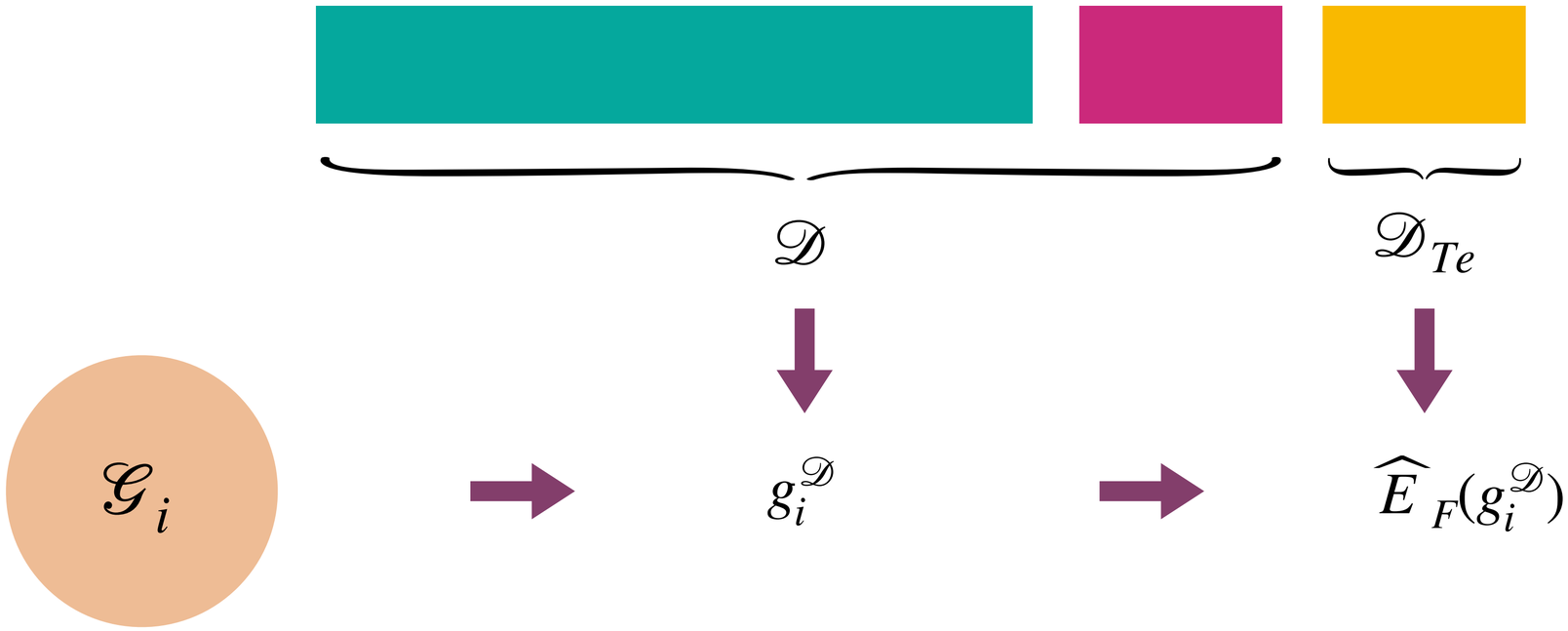
Seleção de modelos



Escolhemos a classe de modelos $\mathscr{G}_i$ que minimiza $\widehat{E}_F(g_i^{\mathscr{D}_{Tr}})$

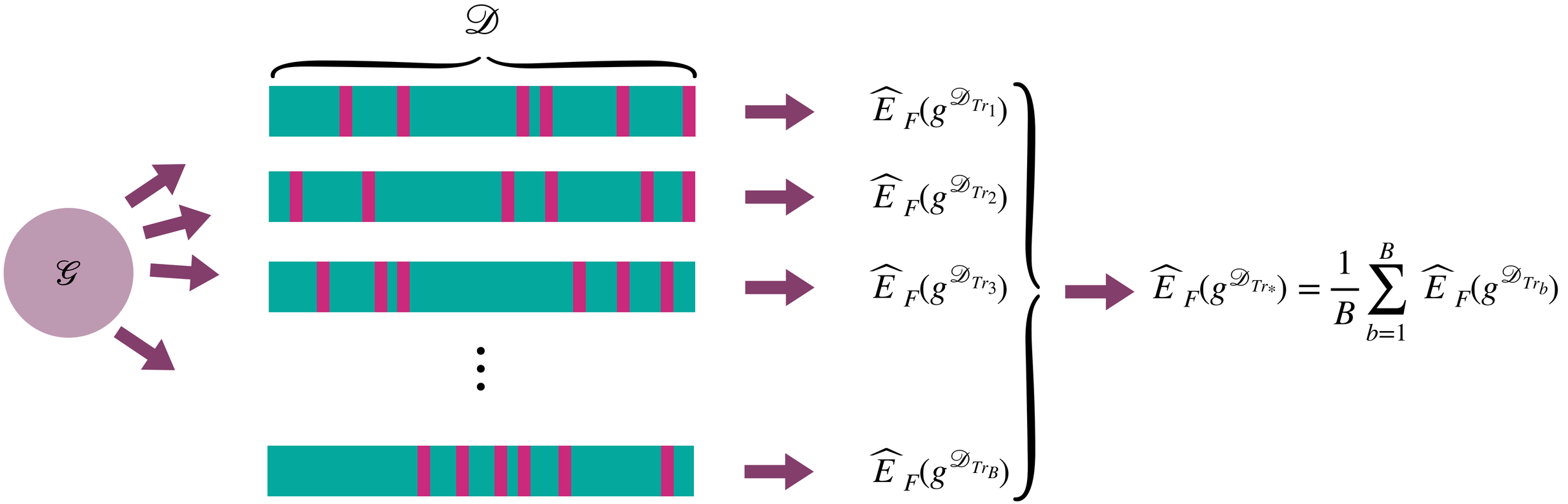
Seleção de modelos

Se $\mathscr{G}_i$ é a classe que minimiza $\widehat{E}_F(g_i^{\mathscr{D}_{Tr}})$



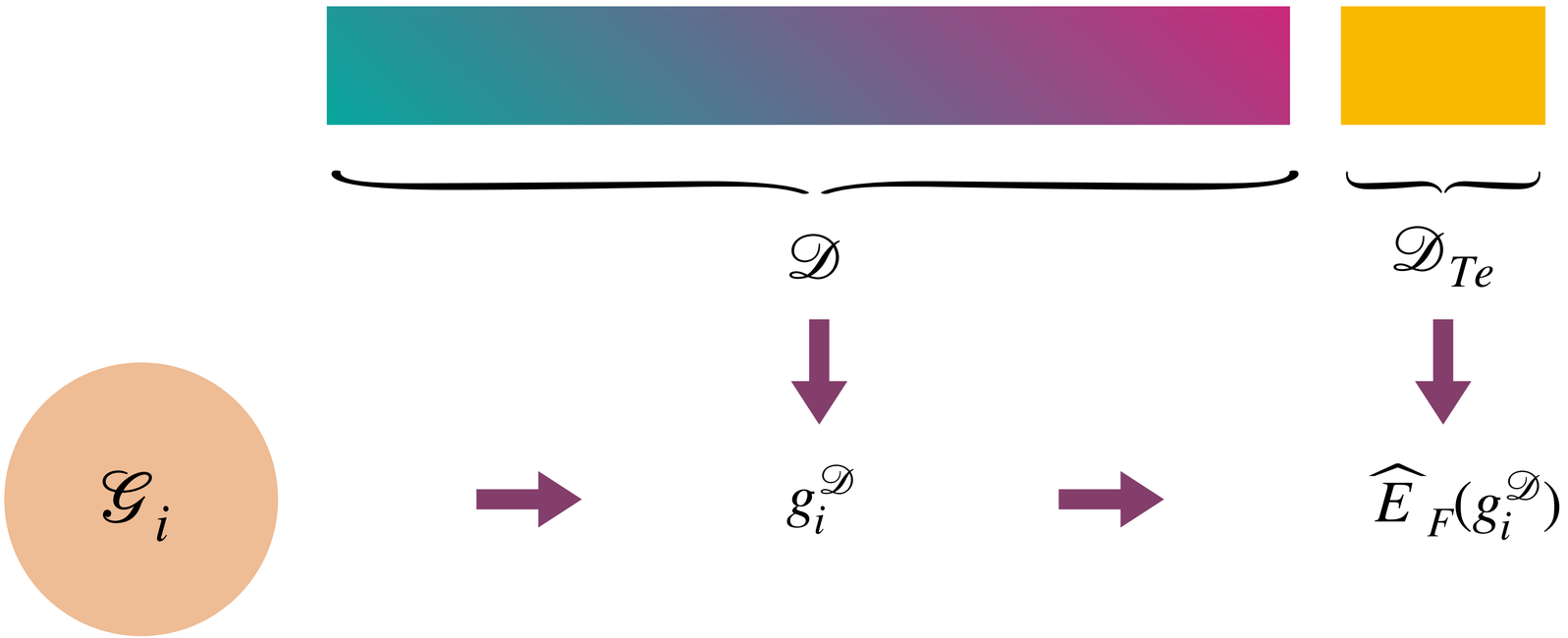
Seleção de modelos por reamostragem (*bootstrap*)

O problema com o esquema treinamento/validação/teste é que temos pouca variabilidade na amostra de treinamento e validação (uma única amostra)



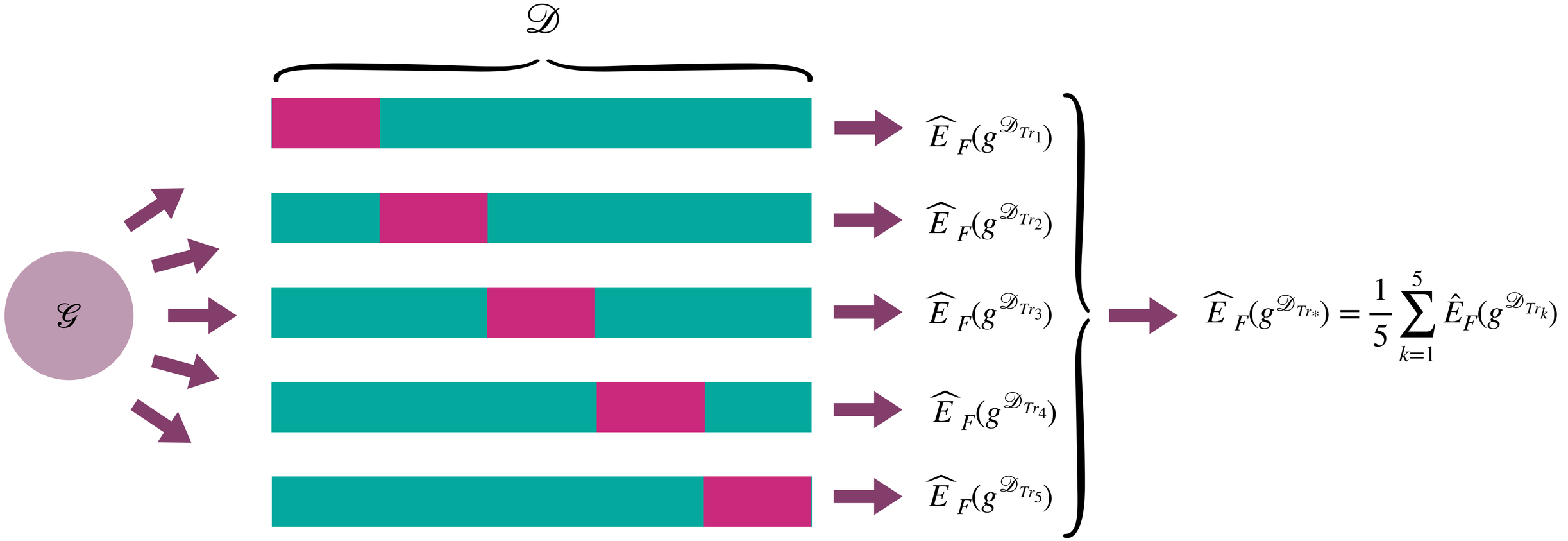
Seleção de modelos por reamostragem (*bootstrap*)

Se $\mathcal{G}_i$ é a classe que minimiza $\widehat{E}_F(g_i^{\mathcal{D}_{Tr*}})$



Seleção de modelos por validação cruzada

Divisão da amostra em k lotes do mesmo tamanho



Em geral $k = 5$ ou $k = 10$

Seleção de modelos por validação cruzada

Se $\mathcal{G}_i$ é a classe que minimiza $\widehat{E}_F(g_i^{\mathcal{D}_{Tr*}})$