Elementos de um problema de predição

INFORMAÇÃO,

TECNOLOGIA

& INOVAÇÃO

REGRESSÃO

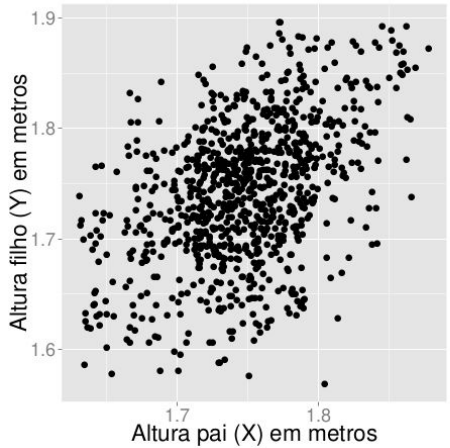
Aprendizado supervisionado: Dadas medições $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, aprender um modelo para prever Y_i baseado em $\mathbf{X}_i$

Y: variável numérica

EXEMPLO

Prever a altura de um filho (Y) com base na altura de seu pai (X)

Amostra: $(X_1, Y_1), \ldots, (X_n, Y_n)$.



Função de predição: $g(x)$

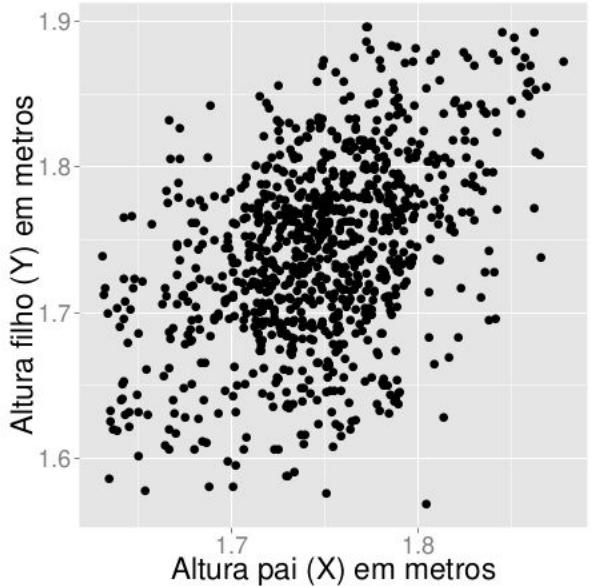
OBJETIVOS

(i) construir g de modo a se obter boas predições

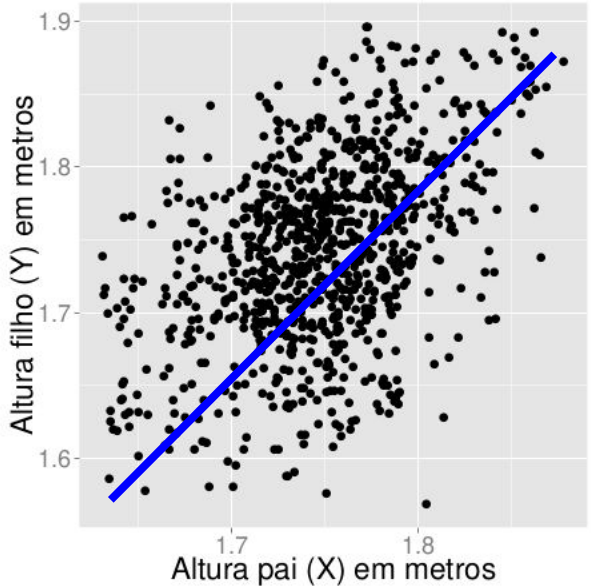
$$g(\mathbf{X}_{n+1}) \approx Y_{n+1}, \ldots, g(\mathbf{X}_{n+m}) \approx Y_{n+m}$$

(ii) saber quantificar o quão boa uma (função de) predição é.

COMO CONSTRUIR g?

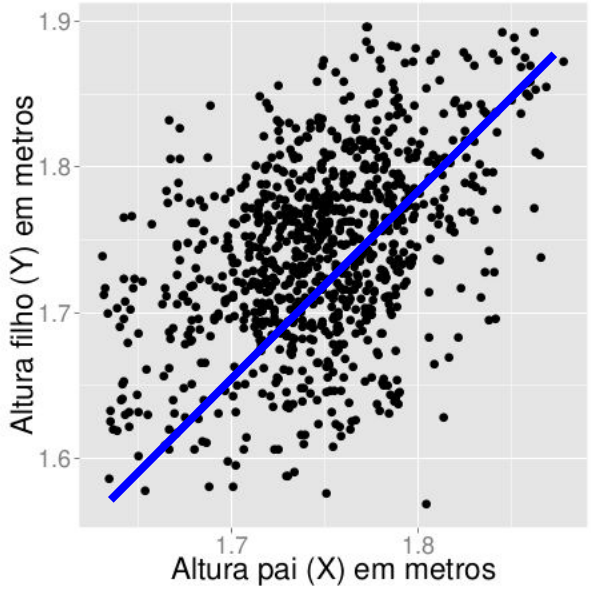


COMO CONSTRUIR g?



Regressão Linear

COMO CONSTRUIR g?



$$x = 1,80\text{m} \Rightarrow g(1,80) = 1,77$$

COMO AVALIAR g?

Se y=1,76m, qual o erro cometido?

Erro quadrático: $(g(x) - y)^2 = (1,77 - 1,76)^2 = 0.0001$

Quanto maior o valor de $(g(x) - y)^2$, pior é nossa predição

COMO AVALIAR g?

Quantificamos quão boa g é apenas para um par (x, y)

Generalização: função de risco

$$R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$$

COMO AVALIAR g?

Se os novos dados seguem a mesma distribuição,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(\mathbf{X}_{n+i}))^2 \approx \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] := R(g)$$

RESUMINDO

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.

  $\mathbf{X} \in \Re^d$ são chamados de preditores, variáveis explicativas, variáveis independentes, covariáveis ou *features*/atributos.

  Y é chamado de resposta, variável dependente ou *labels*

- Desejamos criar uma função de predição $g(\mathbf{x})$ para prever novas observações $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+m}$ bem

- Prever novas observações bem = criar g tal que $R(g)$ seja baixo

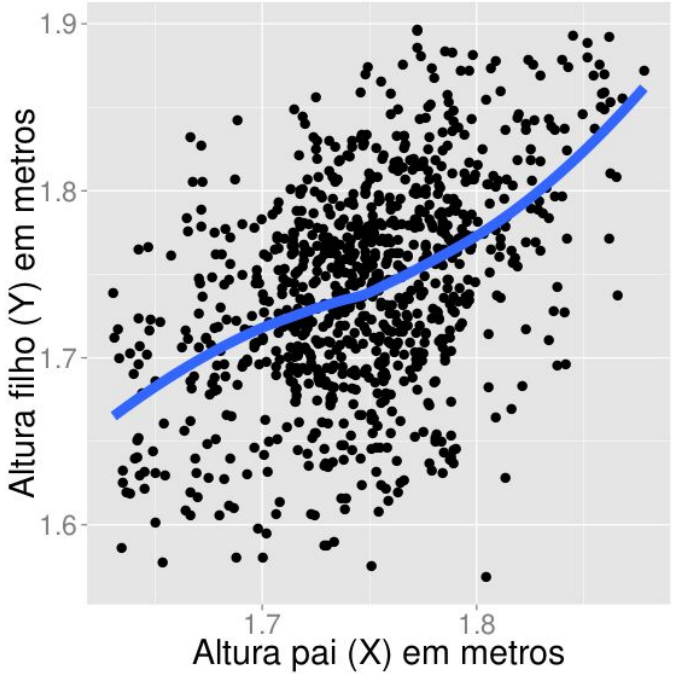
A MELHOR FUNÇÃO DE PREDIÇÃO

Seja $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a função de regressão

Resultado:

$R(r) \leq R(g)$ para toda função $g(\mathbf{x})$

COMO ESTIMAR A FUNÇÃO DE REGRESSÃO?



COMO ESTIMAR A FUNÇÃO DE REGRESSÃO?

Covariáveis				Resposta
$X_{1,1}$	$\dots$	$X_{1,d}$	$(= \mathbf{X}_1)$	Y_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$X_{n,1}$	$\dots$	$X_{n,d}$	$(= \mathbf{X}_n)$	Y_n

Objetivo: estimar $r(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$

$x_{i,j}$: valor da j-ésima covariável no i-ésimo indivíduo.