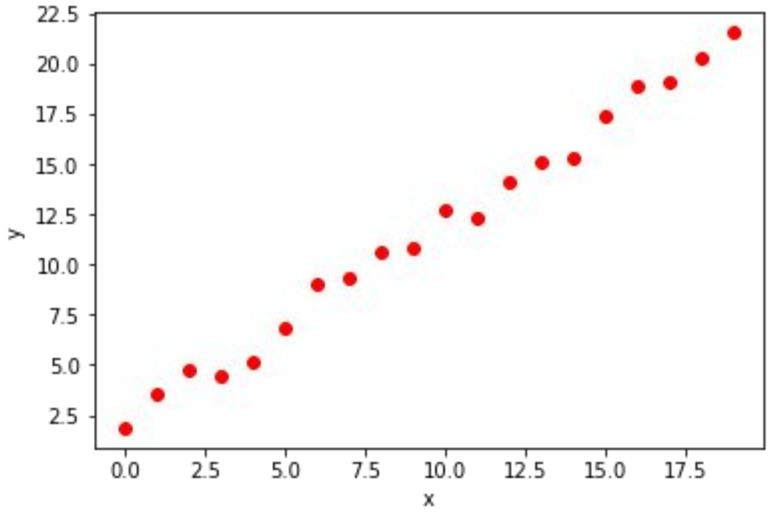
Ajuste de Curvas (e em particular, Regressão Linear)

Fabricio Murai

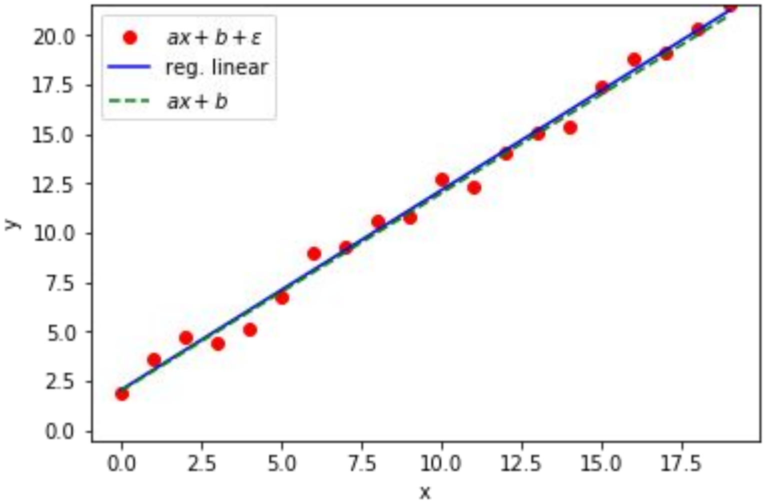
Exemplo de ajuste de curva

- Quero encontrar uma função que relaciona x e y
- A função deve capturar a tendência, mas não precisa passar necessariamente pelos pontos
- Temos que assumir uma "cara" para essa função

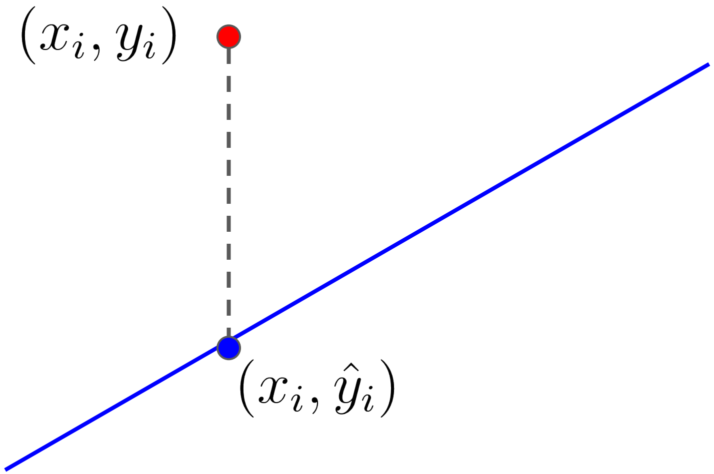


Exemplo de ajuste de curva

- Assuma, por exemplo, f(x) = ax+b (poderíamos ter escolhido outra)
- Queremos encontrar parâmetros a e b que **melhor** ajustam a curva aos dados
- Como definir **melhor** ajuste?



Como medir a qualidade em um ponto x?



Erro absoluto:

$$d_i = |\hat{y}_i - y_i|$$

Erro quadrático:

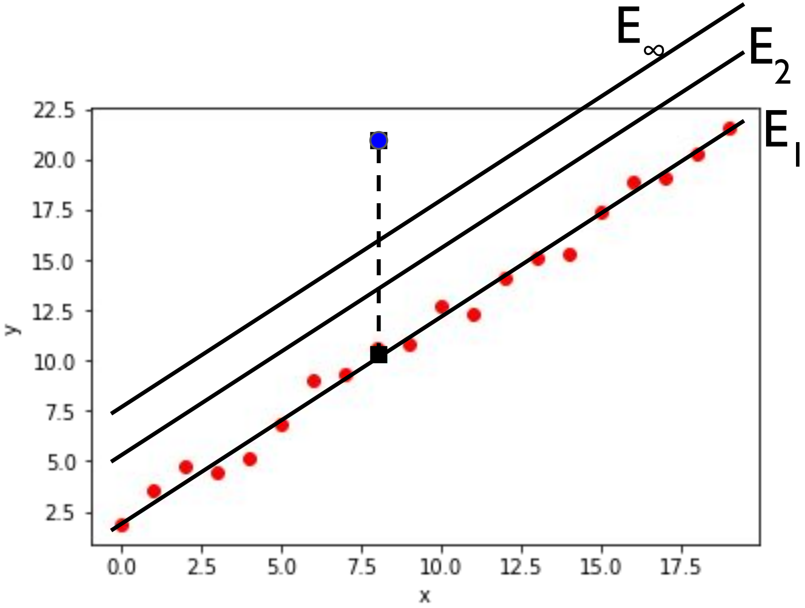
$$d_i = (\hat{y}_i - y_i)^2$$

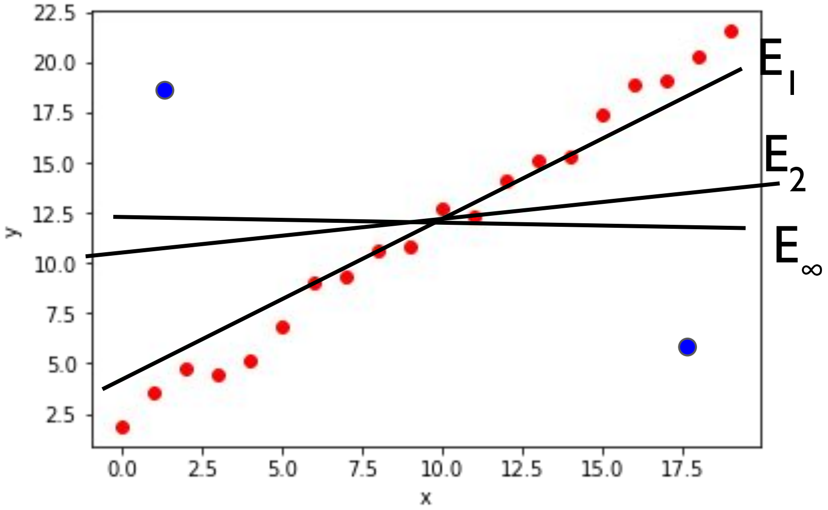
Qualidade do ajuste

Exemplos de métricas levando em consideração pontos (x_1,y_1), …, (x_n,y_n):

1. Erro máximo $$E_\infty = \max_{i=1,\ldots,n} |\hat{y}_i - y_i|$$
2. Erro médio (Erro L_1) $$E_1 = \frac{1}{n} \sum_{i=1,\ldots,n} |\hat{y}_i - y_i|$$
3. Raiz do erro quadrático médio (Erro L_2) $$E_2 = \sqrt{\frac{1}{n} \sum_{i=1,\ldots,n} (\hat{y}_i - y_i)^2}$$

Relação entre métrica de erro e outliers (pontos "fora da curva")





Justificativa para o uso de E_2

Embora o erro médio absoluto E_1 seja mais robusto a outliers, a função módulo não possui derivada contínua.

Já (raiz do) erro médio quadrático é um bom compromisso: relativamente robusto e possui derivada contínua.

Ajuste de curvas visando minimizar (raiz do) erro médio quadrático é conhecido por **Método dos Quadrados Mínimos.**

Método dos Quadrados Mínimos (MQM)

O MQM consiste em encontrar os parâmetros da função f que minimizam o erro

$$E_2 = \sqrt{\frac{1}{n} \sum_{i=1,\dots,n} (\hat{y}_i - y_i)^2}$$

Isto é equivalente a minimizar

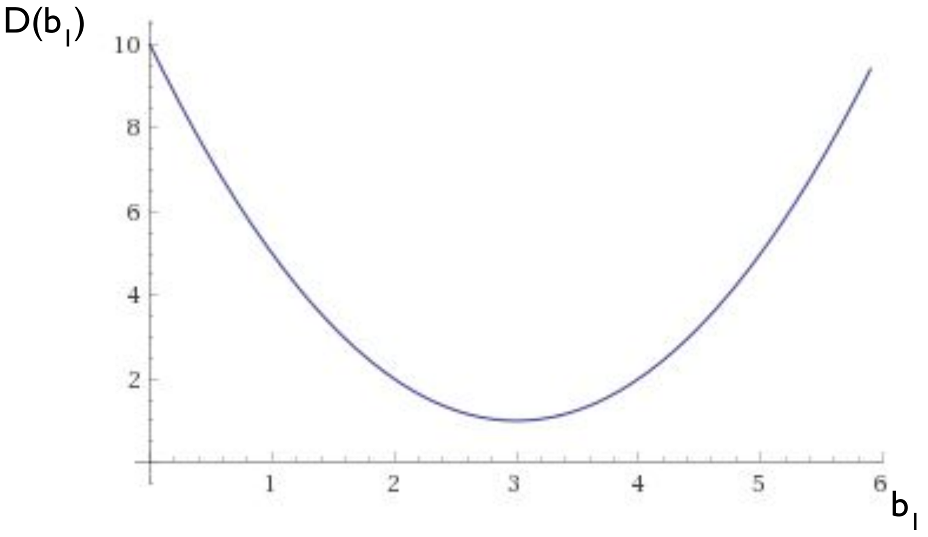
$$D(f) = \sum_{i=1,\dots,n} (\hat{y}_i - y_i)^2$$

Método dos quadrados mínimos

Exemplo 1: $f(x) = b_1 x$

$$D(b_1) = \sum_{k=1}^{n} (f(x_k) - y_k)^2$$
$$= \sum_{k=1}^{n} (b_1 x_k - y_k)^2$$
$$= \sum_{k=1}^{n} \left[b_1^2 x_k^2 - 2b_1 x_k y_k + y_k^2\right]$$
$$= b_1^2 \sum_{k=1}^{n} x_k^2 - 2b_1 \sum_{k=1}^{n} x_k y_k + \sum_{k=1}^{n} y_k^2$$

Qual a "cara" desta função?



Método dos quadrados mínimos

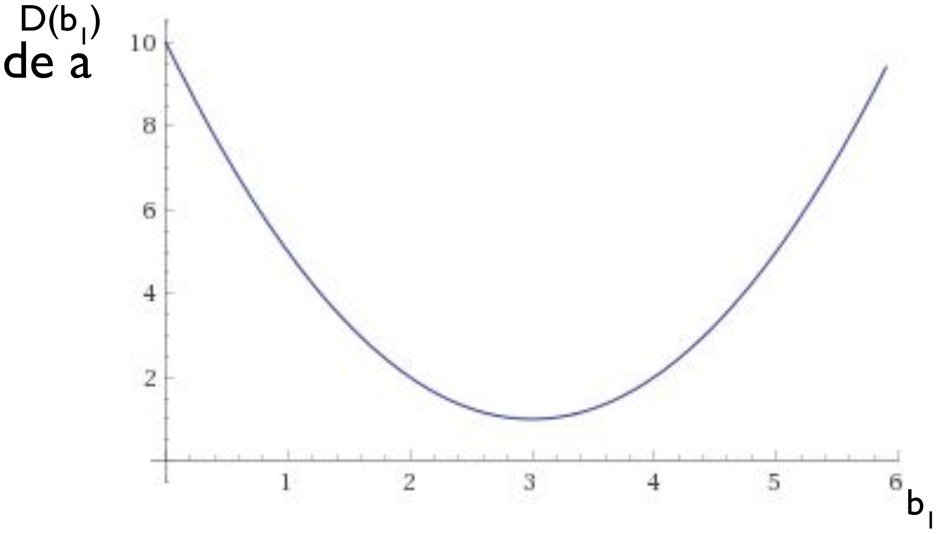
Exemplo 1: $f(x) = b_1x$

A função que queremos minimizar é $D(b_1) = b_1^2 \sum_{k=1}^{n} x_k^2 - 2b_1 \sum_{k=1}^{n} x_k y_k + \sum_{k=1}^{n} y_k^2$

Para isso, queremos encontrar b_1 onde a a derivada é nula.

$$\frac{dD}{db_1} = 2b_1 \sum_{k=1}^{n} x_k^2 - 2 \sum_{k=1}^{n} x_k y_k = 0$$

$$\Rightarrow b_1 = \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} x_k^2}$$



Método dos quadrados mínimos

Exemplo 2: $f(x) = b_0 + b_1 x$

min $D(b_0,b_1)$, onde

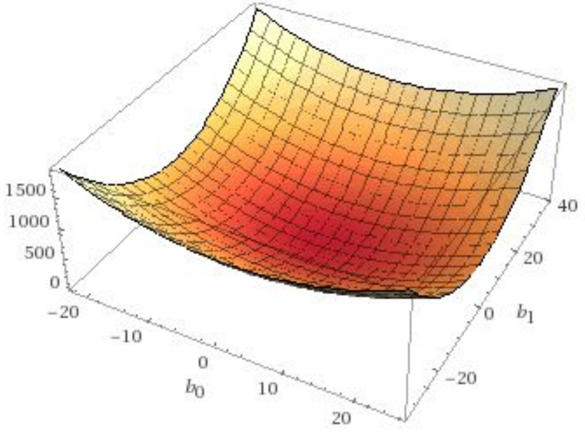
$$D(b_0, b_1) = \sum_{k=1}^{n} (f(x_k) - y_k)^2 = \sum_{k=1}^{n} (b_0 + b_1 x_k - y_k)^2$$

$$\frac{dD}{db_0} = \sum_{k=1}^{n} 2(b_0 + b_1 x_k - y_k) = 0$$

$$\Rightarrow b_0 \sum_{k=1}^{n} 1 + b_1 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k$$

$$\frac{dD}{db_1} = \sum_{k=1}^{n} 2(b_0 + b_1 x_k - y_k) x_k = 0$$

$$\Rightarrow b_0 \sum_{k=1}^{n} x_k + b_1 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} y_k x_k$$



Método dos quadrados mínimos

Exemplo 2: $f(x) = b_0 + b_1 x$

$$b_0 n + b_1 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k$$

$$b_0 \sum_{k=1}^{n} x_k + b_1 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} y_k x_k$$

Como encontrar b_0 e b_1?

Método dos quadrados mínimos

Exemplo 2: $f(x) = b_0 + b_1x$

Como encontrar b_0 e b_1?

$$b_0 n + b_1 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k$$

$$b_0 \sum_{k=1}^{n} x_k + b_1 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} y_k x_k$$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Método dos quadrados mínimos

Exemplo 3: $f(x) = b_0 e^{b_1 x}$

Dá origem a um sistema de equações não-lineares

$$D(b_0, b_1) = \sum_{i=1}^{n} (b_0 e^{b_1 x_i} - y_i)^2$$

$$\frac{\partial D}{\partial b_0} = \sum_{i=1}^{n} 2(b_0 e^{b_1 x} - y_i) e^{b_1 x} = 0$$

$$\frac{\partial D}{\partial b_1} = \sum_{i=1}^{n} 2(b_0 e^{b_1 x} - y_i) b_0 b_1 e^{b_1 x} = 0$$

Generalizando

Relacionamos $x \in R$ com $y \in R$ através de $f(x)$. Chamamos:

- x de **preditor** ou variável independente
- y de **resposta** ou variável dependente

No caso geral, podemos ter um conjunto de preditores representado por um vetor $\mathbf{x}_i=[x_{i1},x_{i2},\ldots,x_{in}]$. Assim, $f(x)$: $R^n \rightarrow R$.

Exemplo: imóveis $(\mathbf{x}_i,y_i)$, onde x_i = [num_quartos$_i$, area$_i$], y = preço$_i$

$f(x) = b_0 + b_1x_{i1}+b_2x_{i2}$

Método dos quadrados mínimos

Podemos escolher outras funções também. Quando estas funções são lineares nos parâmetros b_0, b_1, …, b_p, o MQM dá origem a um sistema de equações lineares.

Quiz: funções lineares nos parâmetros?