Aprendizagem estatística em altas dimensões

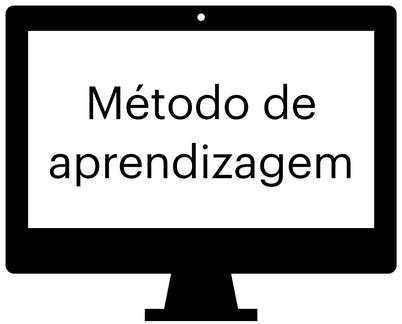
Florencia Leonardi

Conteúdo

- Formulação do problema de aprendizagem estatística, função objetivo, função de custo
- Diferentes tipos de erro
- Modelo linear para regressão
- Estimador de mínimos quadrados
- Transformação de variáveis

Revisão da aula anterior

$$\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$



$$g$$

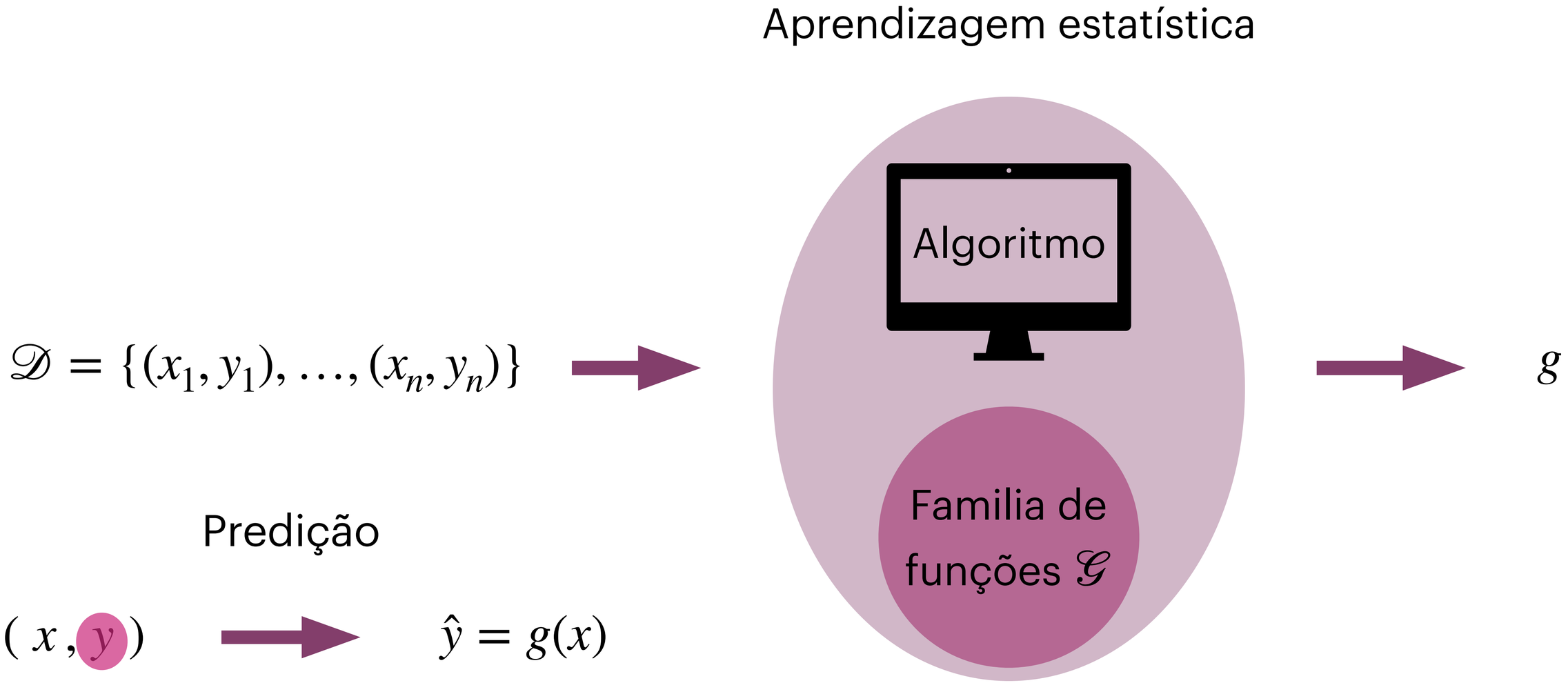
Predição

$$(\, x \,, y \,) \qquad \hat{y} = g(x) \qquad \hat{y} \approx y \ ?$$

O objetivo é escolher g de tal forma que a predição $\hat{y}$ esteja "próxima" de y

Como y e $\hat{y}$ são variáveis aleatórias, buscamos minimizar $\mathbb{E}(L(y, \hat{y}))$ para alguma função de custo L escolhida antes da análise dos dados e de forma adequada para o problema

Métodos de aprendizagem estatística



Principais desafios da aprendizagem estatística supervisionada

Dada a função de custo L e a amostra $\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$:

- Como estimar $\mathbb{E}[L(y, g(x))]$ para uma função $g \in \mathscr{G}$ escolhida com base em $\mathscr{D}$?
- Como escolher g de forma a minimizar $\mathbb{E}[L(y, g(x))]$?

→ Estes são os principais desafios na aprendizagem estatística supervisionada

Formalização do problema

- (X, Y) variáveis aleatórias com densidade conjunta $p(x, y)$, com valores em $\mathcal{X} \times \mathcal{Y}$
- Uma função objetivo $f\colon \mathcal{X} \to \mathcal{Y}$ desconhecida

Assumimos que X e Y estão relacionadas por meio da função objetivo f através da equação

$$Y = f(X) + \epsilon$$

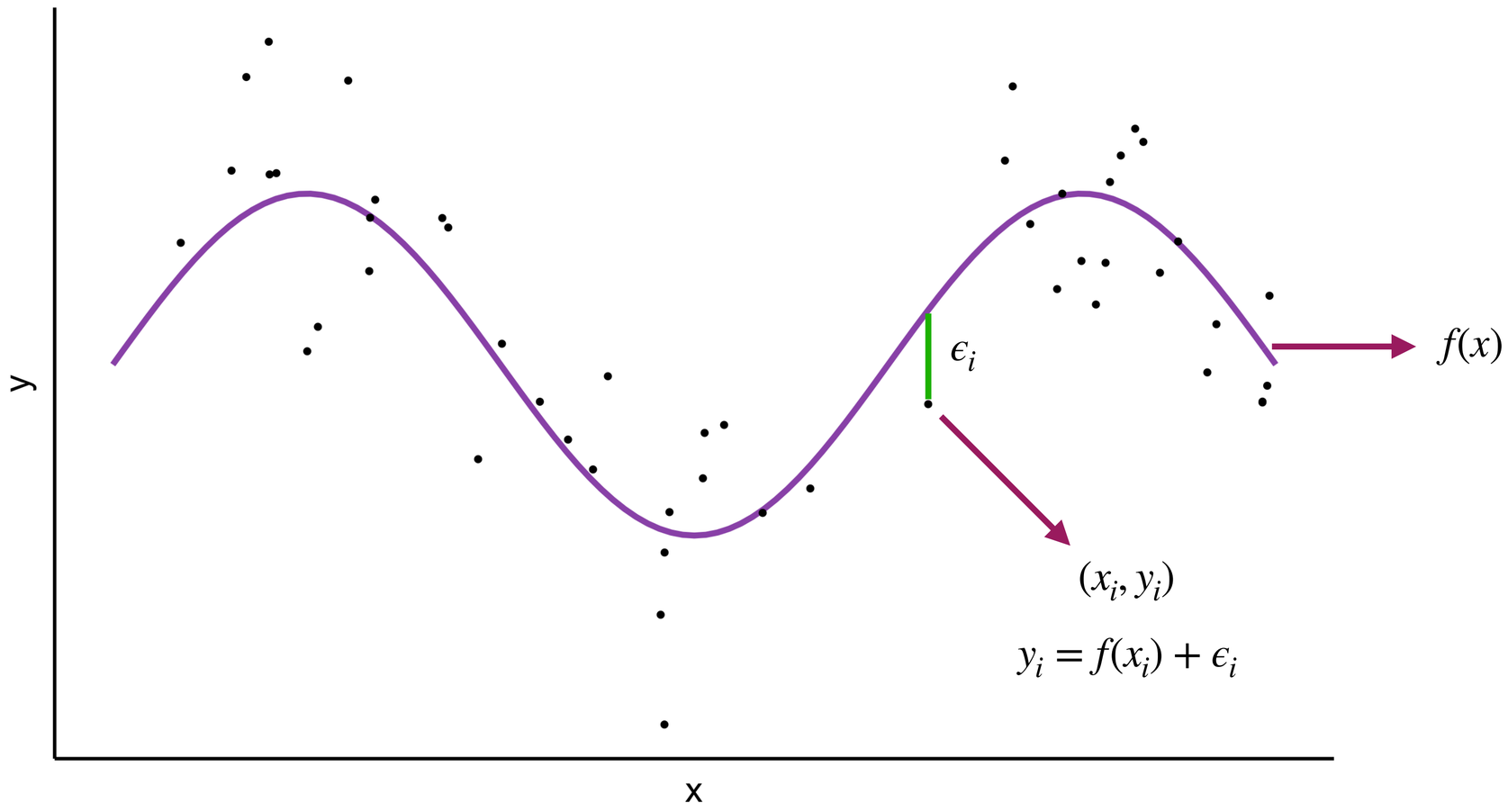
onde ϵ é uma variável aleatória com $\mathbb{E}(\epsilon) = 0$ e $\text{Var}(\epsilon) = \sigma^2$

Formalização do problema

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

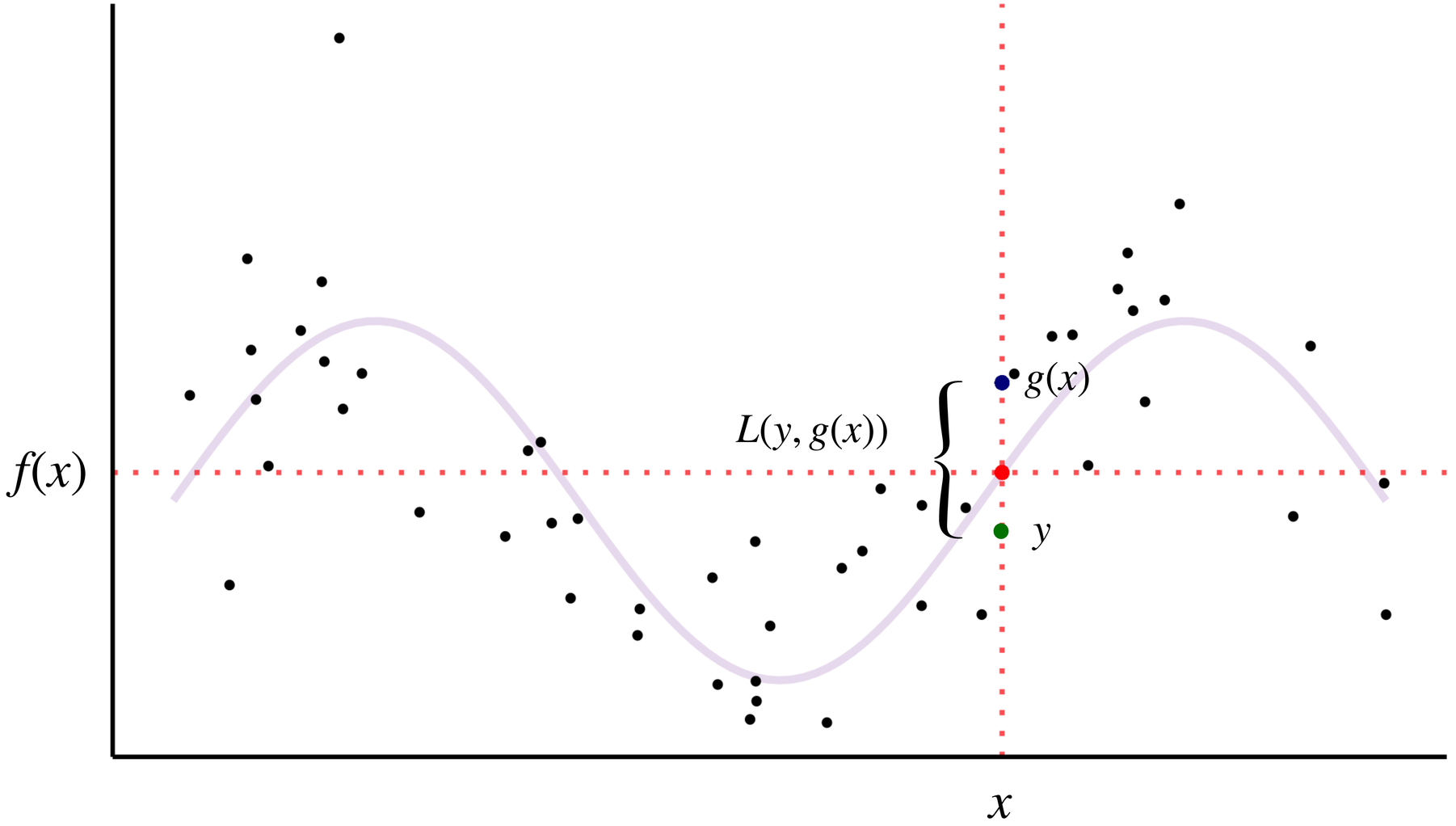
- Assume-se que $\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ tem a mesma distribuição de (X, Y)
- A busca por uma função que “aproxime" f é feita numa família de funções candidatas $\mathscr{G}$
- A maioria dos métodos de aprendizagem tentam encontrar a função $g \in \mathscr{G}$ que minimize $\mathbb{E}[L(y, g(x))]$, para uma função de custo $L\colon \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_{\geq 0}$ previamente definida

Formalização do problema



$\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ com distribuição $p(x, y)$

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 !

Modelo linear para regressão

Diferentes modelos e métodos especificam uma forma diferente para a escolha da função $g \in \mathcal{G}$ (lembremos que f sempre é desconhecida!)

No modelo de regressão linear, a família $\mathcal{G}$ é uma família de funções lineares:

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

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

$$x^T\beta \ = \ \underbrace{(x_1 \ x_2 \ \dots \ x_p)}_{\mathbb{R}^{1\times p}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}}_{\mathbb{R}^{p\times 1}} = \underbrace{\sum_{j=1}^{p} x_j\beta_j}_{\mathbb{R}^{1\times 1}}$$

Modelo linear para regressão

Por uma questão de simplicidade na notação vamos fazer a identificação $x \mapsto (1, x)$ assim o modelo fica descrito como

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

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \qquad x^T\beta = \beta_0 + \sum_{j=1}^{p} x_j\beta_j$$

Tipos de erro

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

$$\widehat{E}_D(g) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, g(x_i))$$ → 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"

Modelo linear para regressão

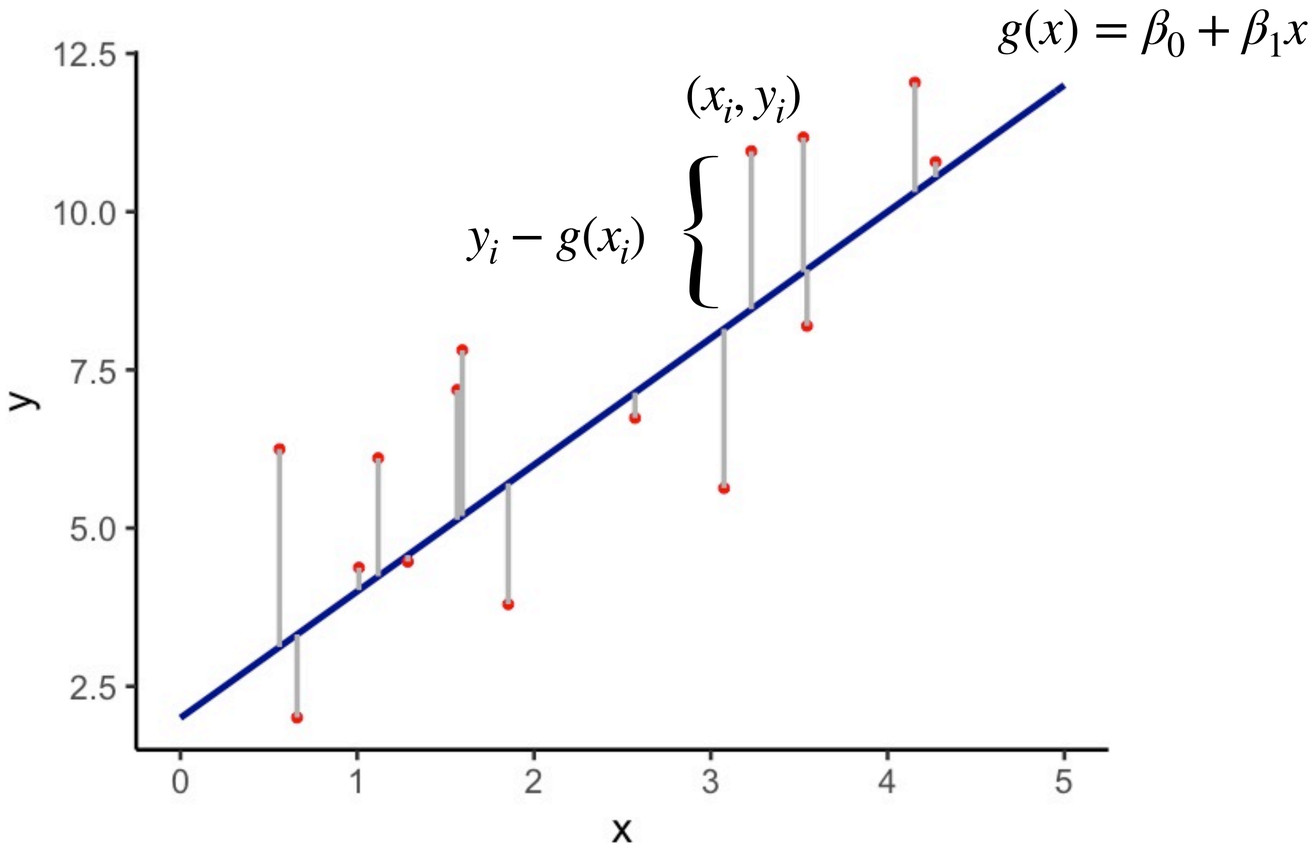
Para os modelos de regressão em geral usamos a função de custo quadrática
$L(y, g(x)) = (y - g(x))^2$

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

Isto é equivalente a escolher $\beta \in \mathbb{R}^{p+1}$ que minimize $\widehat{E}_D(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$

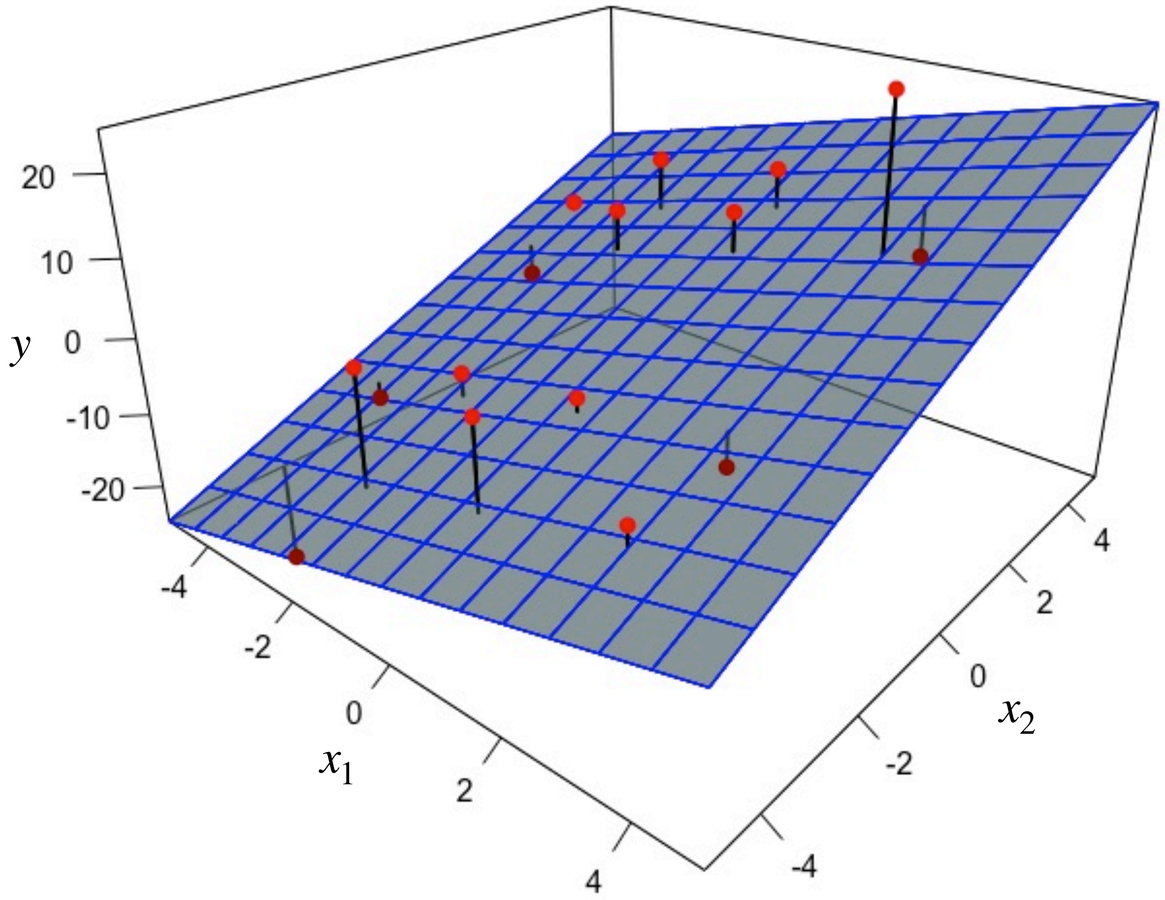
O vetor β obtido pela minimização de $\widehat{E}_D(\beta)$, denotado por $\hat{\beta}$, é conhecido como estimador de mínimos quadrados

Modelo linear para regressão



$$\mathcal{G} = \{g(x) = x^T\beta,\ \ \beta \in \mathbb{R}^2\}$$

Modelo linear para regressão



$$\mathcal{G} = \{g(x) = x^T\beta,\ \ \beta \in \mathbb{R}^3\}$$

Modelo linear para regressão

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

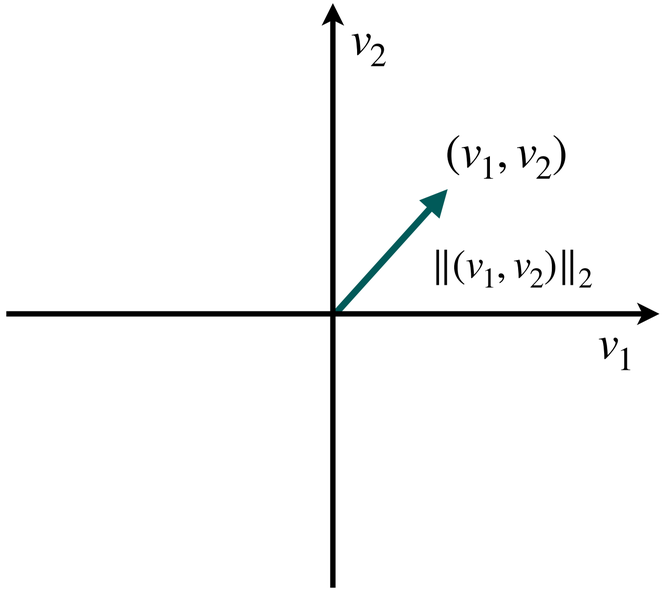
Observemos que podemos escrever $\widehat{E}_D(\beta)$ como $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1(p+1)} \\ 1 & x_{22} & \cdots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \cdots & x_{n(p+1)} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Modelo linear para regressão

$$\widehat{E}_D(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2/n$$

$$\|(v_1, \ldots, v_n)\|_2^2 = \sum_{i=1}^{n} v_i^2$$



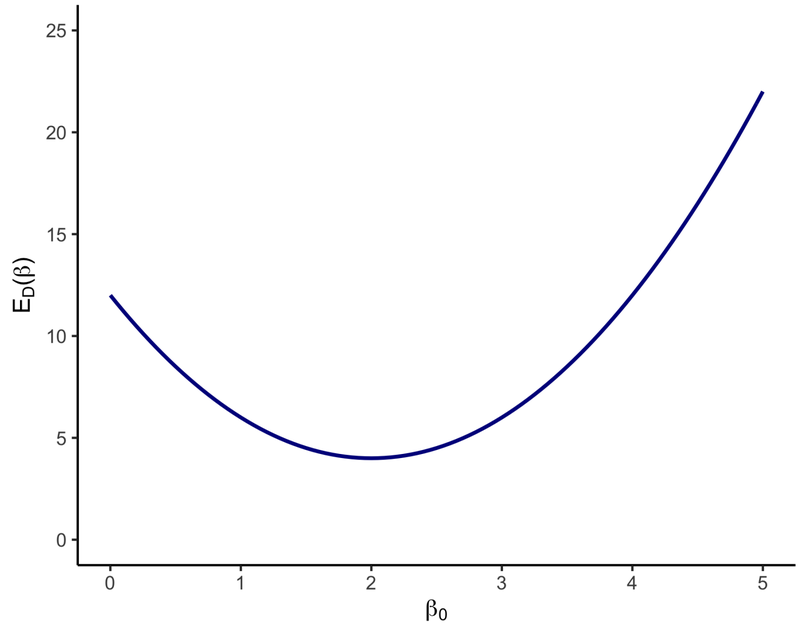
$$\mathbf{y} - \mathbf{X}\beta = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \cdots & x_{1(p+1)} \\ 1 & x_{22} & \cdots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \cdots & x_{n(p+1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} y_1 - x_1^T\beta \\ y_2 - x_2^T\beta \\ \vdots \\ y_n - x_n^T\beta \end{pmatrix}$$

$$\|\mathbf{y} - \mathbf{X}\beta\|_2^2/n = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

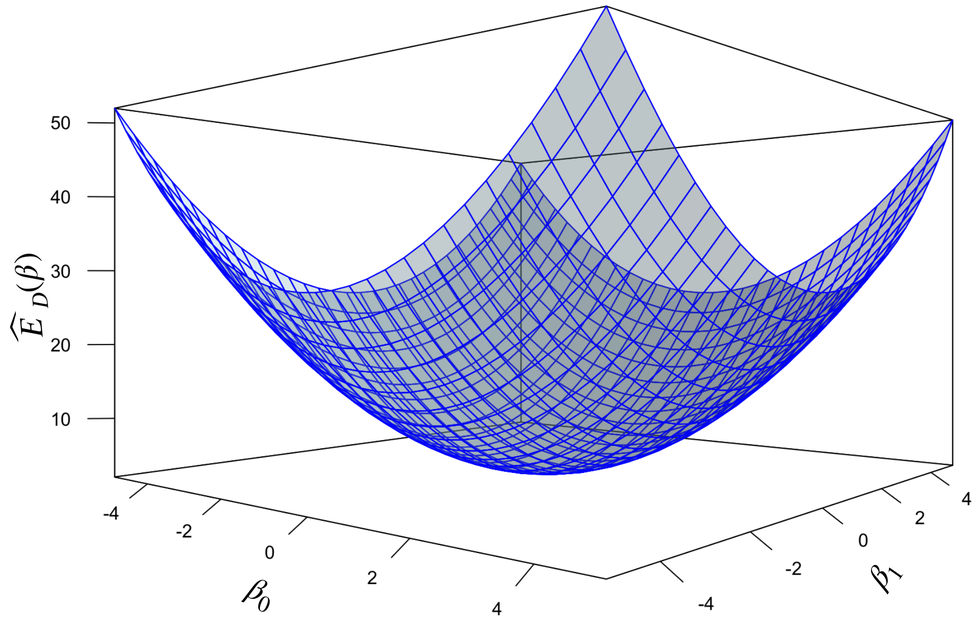
Modelo linear para regressão

Objetivo: escolher $\beta \in \mathbb{R}^{p+1}$ que minimize $\widehat{E}_D(\beta) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$

$\widehat{E}_D(\beta)$ é uma função convexa de β



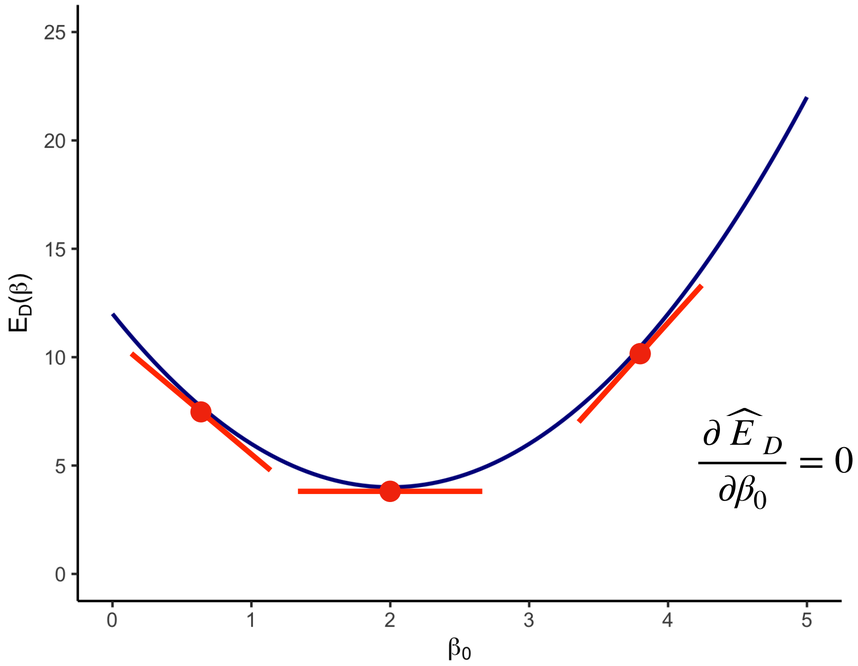
$\beta \in \mathbb{R}$

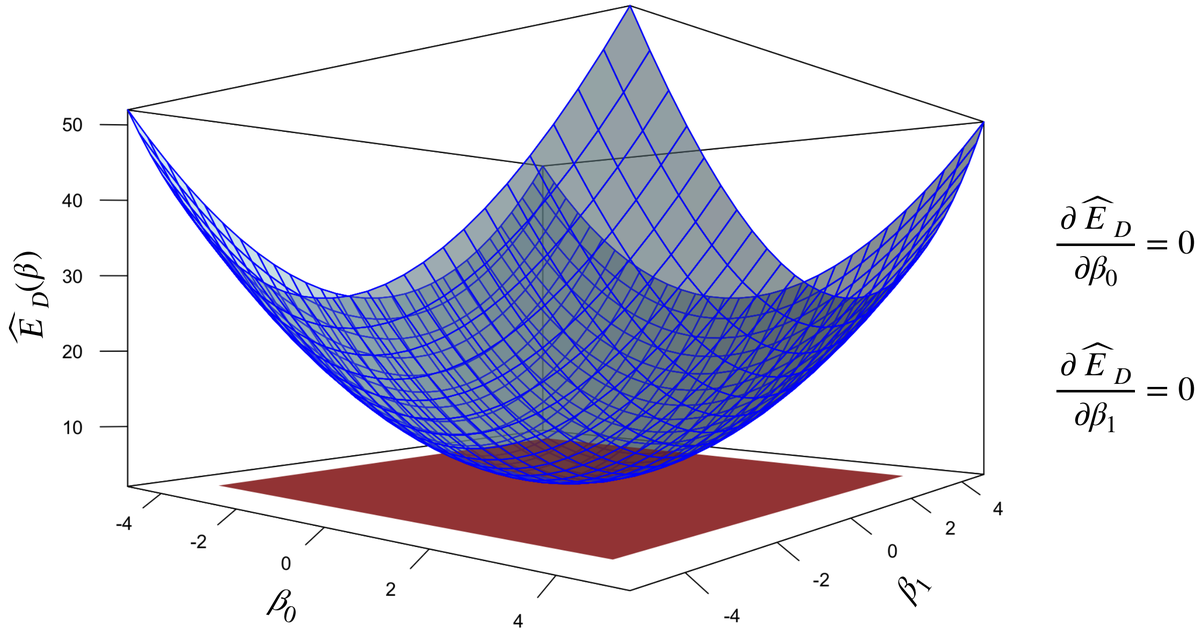


$\beta \in \mathbb{R}^2$

Modelo linear para regressão

Objetivo: escolher $\beta \in \mathbb{R}^{p+1}$ que minimize $\widehat{E}_D(\beta) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$





Modelo linear para regressão

Uma função convexa sempre tem pelo menos um mínimo (o mínimo pode não ser único)

No caso em que $\beta \in \mathbb{R}$ podemos derivar $\widehat{E}_D(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2$ e obtemos que

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

Fazendo $\frac{\partial \widehat{E}_D(\beta)}{\partial \beta} = 0$ obtemos que a solução é $\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

Modelo linear para regressão

$$\widehat{E}_D(\beta) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

No caso geral $\beta \in \mathbb{R}^{p+1}$, para obter $\hat{\beta}$ devemos calcular $\frac{\partial \widehat{E}_D(\beta)}{\partial \beta_i}$ para todo $i = 1,\ldots,p+1$ e encontrar β tal que $\frac{\partial \widehat{E}_D(\beta)}{\partial \beta_i} = 0$ para todo $i = 1,\ldots,p+1$

Se as colunas de $\mathbf{X}$ são linearmente independentes então há uma única solução e está dada por $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1(p+1)} \\ 1 & x_{22} & \cdots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \cdots & x_{n(p+1)} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

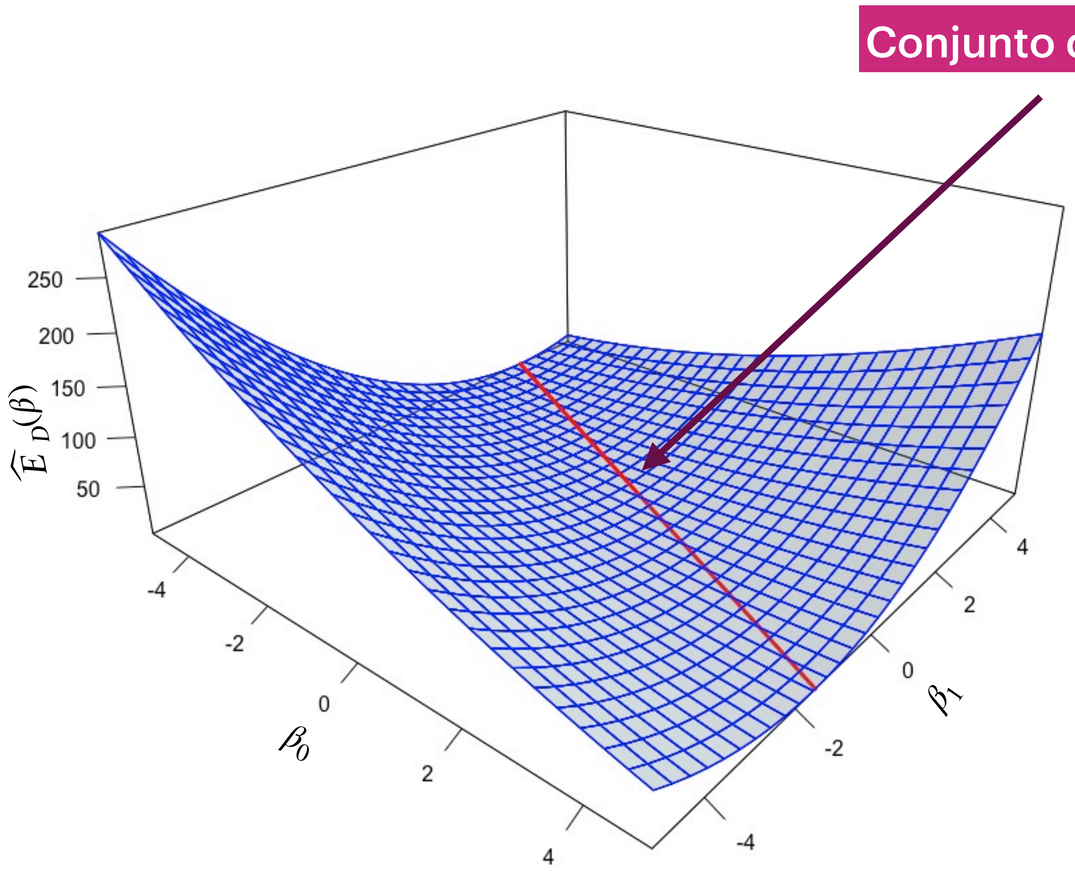
Modelo linear para regressão

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\beta} = \left(\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & & & \\ x_{1(p+1)} & x_{2(p+1)} & \dots & x_{n(p+1)} \end{pmatrix} \begin{pmatrix} 1 & x_{12} & \dots & x_{1(p+1)} \\ 1 & x_{22} & \dots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \dots & x_{n(p+1)} \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{12} & x_{22} & \dots & x_{n2)} \\ \vdots & & & \\ x_{1(p+1)} & x_{2(p+1)} & \dots & x_{n(p+1)} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{(p+1)\times 1}$$

$\mathbb{R}^{(p+1)\times n}$ $\mathbb{R}^{n\times(p+1)}$ $\mathbb{R}^{(p+1)\times n}$ $\mathbb{R}^{n\times 1}$

Modelo linear para regressão

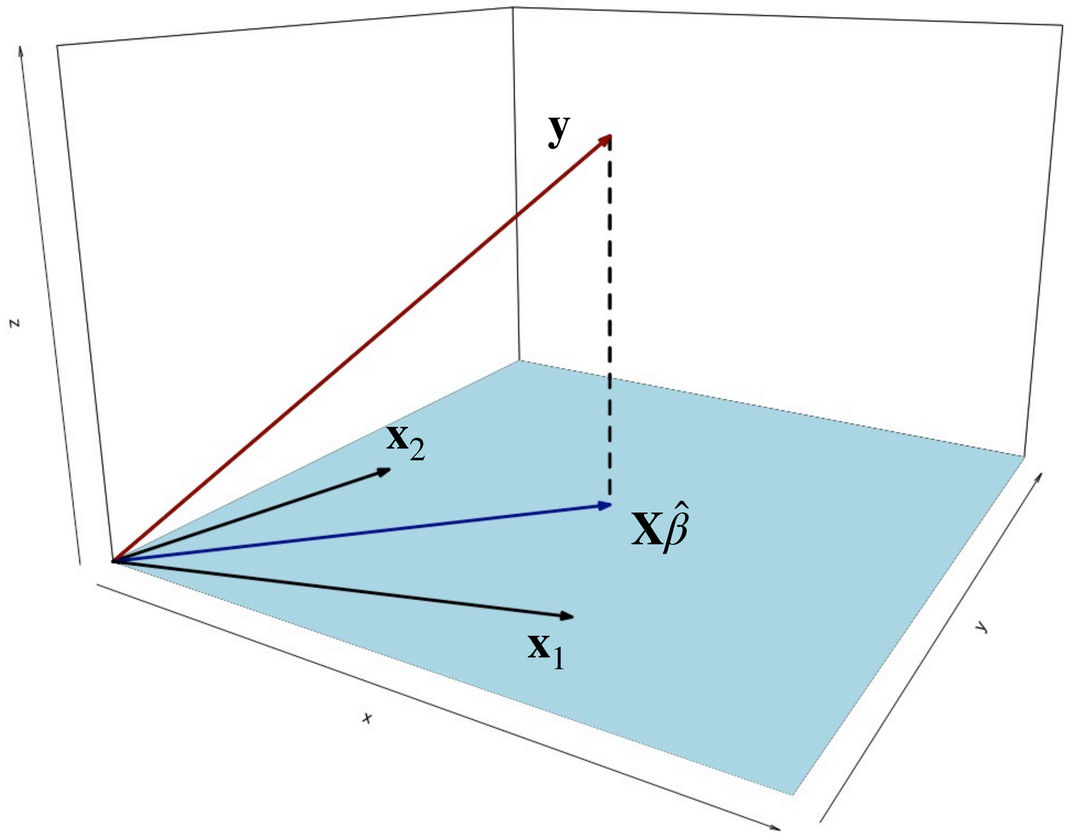


Em alguns casos o problema de minimizar $\widehat{E}_D(\beta)$ não tem uma solução única

Isso acontece quando as colunas de $\mathbf{X}$ são linearmente dependentes e $\mathbf{X}^T\mathbf{X}$ não é invertível

É o caso de dados em alta dimensão: $p \geq n$

Modelo linear para regressão



Exemplo com $n = 3$ e $p = 1$

Projeção do vetor $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

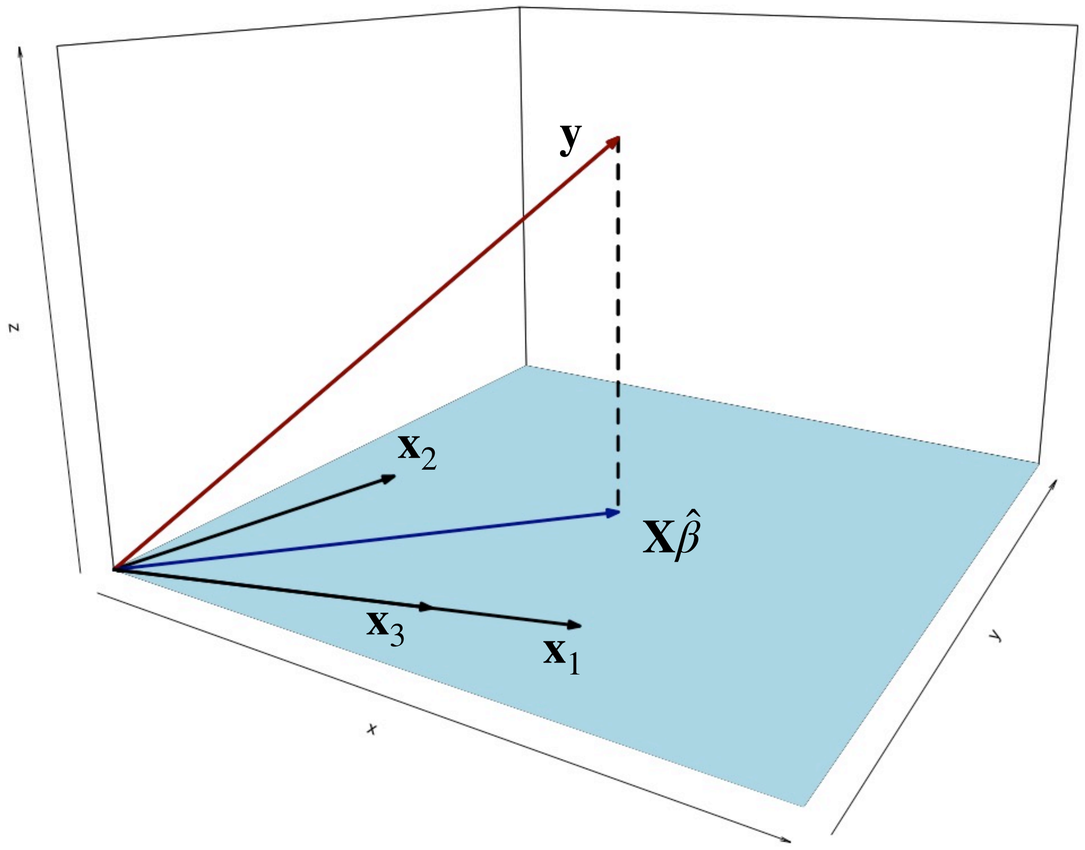
no espaço gerado pelas colunas de

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1(p+1)} \\ 1 & x_{22} & \cdots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \cdots & x_{n(p+1)} \end{pmatrix}$$

dadas por $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$.

Como as colunas de $\mathbf{X}$ são linearmente independentes, o vetor $\hat{\beta}$ é único

Modelo linear para regressão



Exemplo com $n = 3$ e $p = 2$

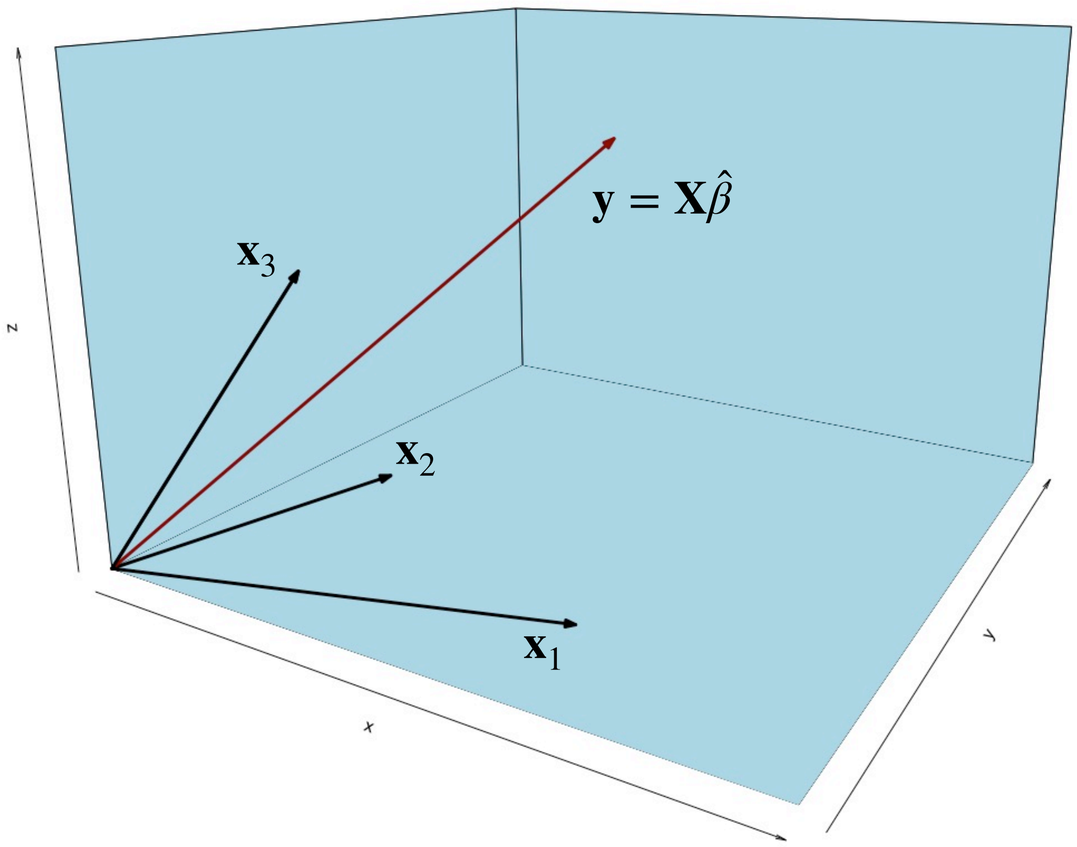
Projeção do vetor $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

no espaço gerado pelas colunas de $\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1(p+1)} \\ 1 & x_{22} & \cdots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \cdots & x_{n(p+1)} \end{pmatrix}$

dadas por $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$.

Como as colunas de $\mathbf{X}$ não são linearmente independentes, o vetor $\hat{\beta}$ não é único

Modelo linear para regressão



Exemplo com $n = 3$ e $p = 2$

Projeção do vetor $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

no espaço gerado pelas colunas de $\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1(p+1)} \\ 1 & x_{22} & \cdots & x_{2(p+1)} \\ \vdots & & & \\ 1 & x_{n2} & \cdots & x_{n(p+1)} \end{pmatrix}$

dadas por $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$.

Como as colunas de $\mathbf{X}$ geram todo o espaço temos que

$\mathbf{y} = \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ e $\widehat{E}_D(\hat{\beta}) = 0$

Exemplo

Price	Condo	Size	Negotiation type
930	220	47	rent
1000	148	45	rent
1000	100	48	rent
990000	870	121	sale
410000	630	51	sale
820000	1000	109	sale

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \begin{pmatrix} x_{12} & \dots & x_{13} \\ x_{22} & \dots & x_{23} \\ \vdots & & \\ x_{n2} & \dots & x_{n3} \end{pmatrix}$$

$$x_{i3} = \begin{cases} 1 & \text{se Negotiation.Type = "rent"} \\ 0 & \text{se Negotiation.Type = "sale"} \end{cases}$$

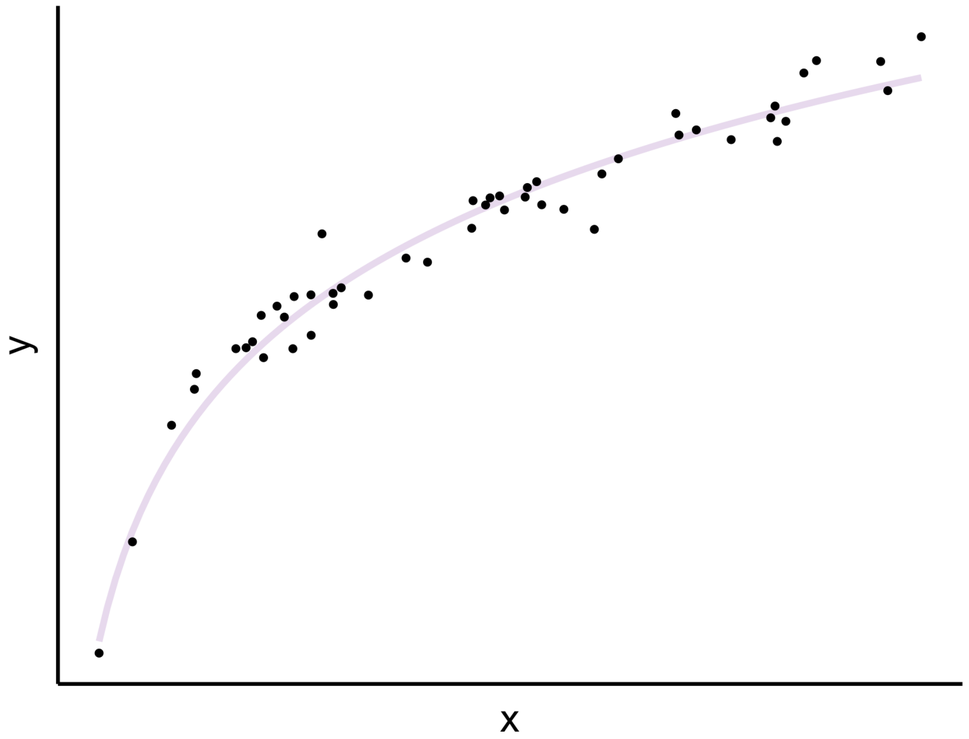
Exemplo

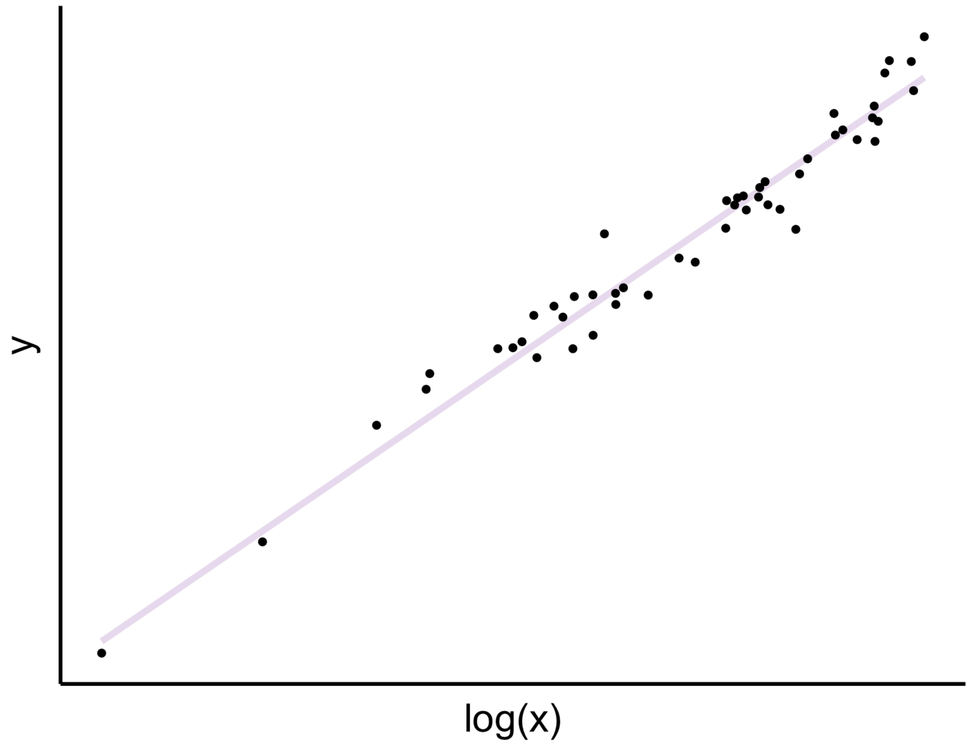
```
dados <- read.csv("sao-paulo-properties-april-2019.csv")

modelo <- lm(Price~Condo+Size+Negotiation.Type, dados)

modelo
```

```
Call:
lm(formula = Price ~ Condo + Size + Negotiation.Type, data = dados)

Coefficients:
         (Intercept)                 Condo                  Size  Negotiation.Typesale
           -387272.2                  -35.6                4690.4                646308.3
```

Transformação de variáveis

Em muitas aplicações y pode não “depender” linearmente de x





Transformação de variáveis

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^{p+1} \quad \longrightarrow \quad \phi(x) = \begin{pmatrix} 1 \\ \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_q(x) \end{pmatrix} \in \mathbb{R}^{q+1}$$

$\Phi = \{1, \phi_1, \ldots, \phi_q\}$ é chamada de “dicionário”

Transformação de variáveis

Neste caso, a família de funções $\mathcal{G}$ é

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

$$\phi(x) = \begin{pmatrix} 1 \\ \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_q(x) \end{pmatrix}$$

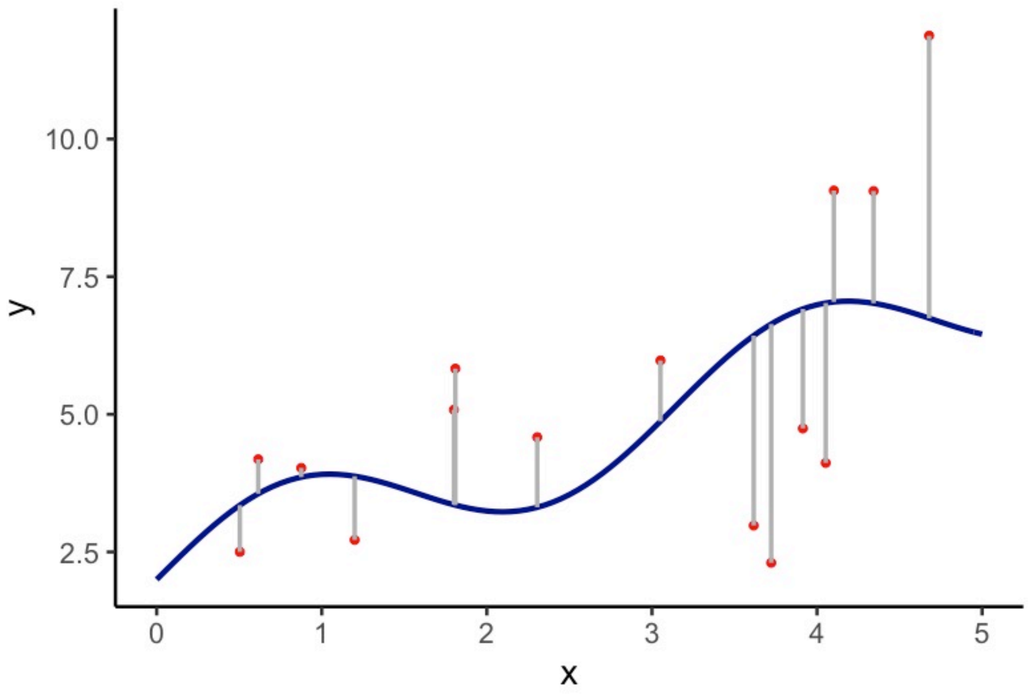
O vetor β que minimiza $\widehat{E}_D(\beta)$ pode ser encontrado da mesma forma que no caso anterior, considerando o vetor $\tilde{x} = \phi(x)$ no lugar de x

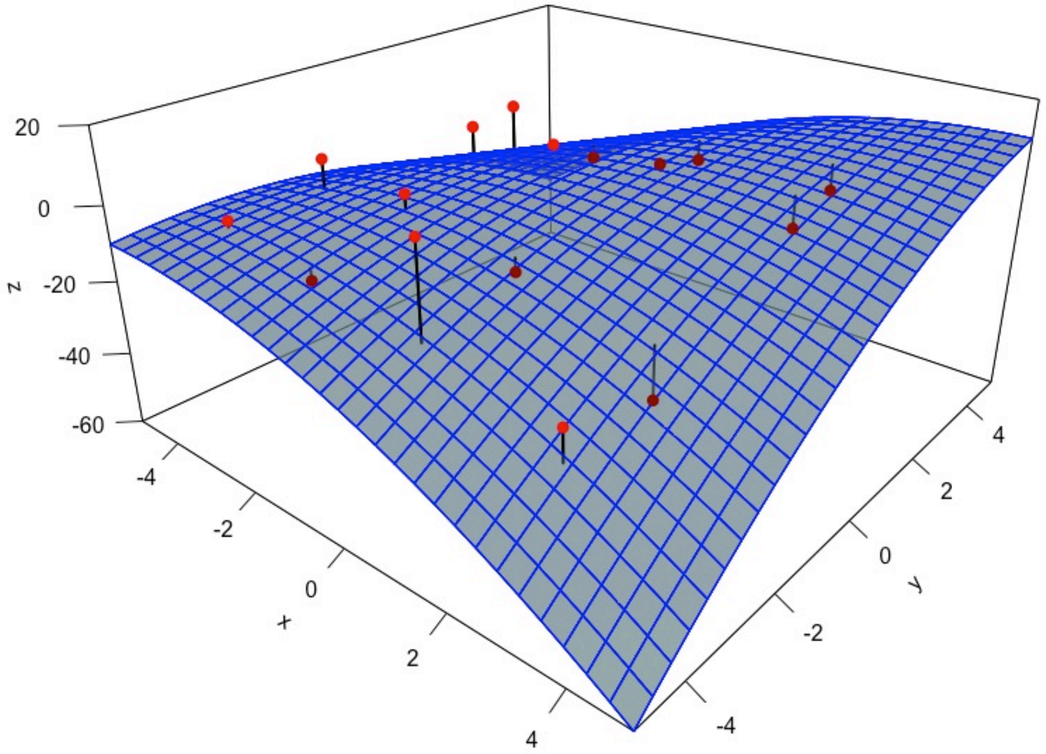
$$\hat{\beta} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$$

$$\tilde{\mathbf{X}} = \begin{pmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{pmatrix}$$

$$\phi(x)^T = \begin{pmatrix} 1 & \phi_1(x) & \phi_2(x) & \dots & \phi_q(x) \end{pmatrix}$$

Transformação de variáveis





Modelo linear para classificação?

- Num problema de classificação temos $y \in \mathcal{Y} = \{c_1, \ldots, c_K\}$ então o modelo linear anterior não é adequado para modelar y diretamente
- Neste caso, podemos modelar como resposta o vetor de probabilidades condicionais $p(y = c_k \,|\, x), \quad k = 1, \ldots, K$ e depois definir, por exemplo $\hat{y} = \arg\max_{c_k} \hat{p}(c_k \,|\, x)$
- Como as probabilidades pertencem ao intervalo $[0,1]$ precisamos uma função que "mapeie" a predição do modelo linear $g(x)$ com este intervalo
- Numa próxima aula veremos como fazer isso...