Least Squares as Orthogonal Projection

Renato Assunção
ESRI Inc. and DCC-UFMG

Dois problemas clássicos em Machine Learning

- ⊕ Regressão
- ⊕ Classificação

⊕ Regressão: predizer Y = preço de ap$^{\underline{to}}$ a partir de características (features) tais como:

X_1 = área do ap$^{\underline{to}}$

X_2 = idade do ap$^{\underline{to}}$

X_3 = n$^{\underline{o}}$ de quartos

⋮

X_{30} = n$^{\underline{o}}$ vagas de garagem

$$Y = \text{preço do ap}^{\text{to}}$$

$$\underset{\sim}{X} = \Big(\underset{\text{área}}{\underset{\downarrow}{X_1}},\ \underset{\text{idade}}{\underset{\downarrow}{X_2}},\ \ldots,\ \underset{\text{vagas}}{\underset{\downarrow}{X_{30}}}\Big) \in \mathbb{R}^{30}$$

Numerical variables

<u>Dados</u>

Amostra

$$\left\{\begin{array}{c|cccc}
Y_1 & X_{11} & X_{12} & \cdots & X_{1;30} \\
Y_2 & X_{21} & X_{22} & \cdots & X_{2,30} \\
Y_3 & X_{31} & X_{32} & \cdots & X_{3,30} \\
\vdots & \vdots & \vdots & & \vdots \\
Y_{2000} & X_{2000,1} & X_{2000,2} & \cdots & X_{2000,30} \\
\downarrow & \downarrow & \downarrow & & \downarrow \\
\text{preços} & \text{área} & \text{idade} & & \text{vagas}
\end{array}\right.$$

③

Queremos encontrar uma fórmula (função)

$$g(\underset{\sim}{X}) = g(\text{área}, \text{idade}, \ldots, \text{vagas})$$

tal que $|Y - g(\underset{\sim}{X})| \simeq 0$

para todo elemento da amostra (todo apto)

⊕ $Y - g(\underset{\sim}{X})$ = erro ao predizer preço Y usando $g(\underset{\sim}{X})$ como preditor

- O que eu <u>realmente</u> quero é predizer preços de ap^{to} que eu não observei ainda

- <u>Conceito</u>:
  - selecione um novo ap^{to} ao acaso $(Y, \underset{\sim}{X}) = (Y, x_1, \ldots, x_{30})$
  - Qual o <u>valor esperado</u> do erro de predição $\mathbb{E}\left(|Y - g(\underset{\sim}{X})|\right)$

⊕ Resultados práticos e teóricos para o erro ao quadrado:

$$MSE = \mathbb{E}\left[\underbrace{\left(Y - g(\underset{\sim}{X})\right)^2}_{\text{erro de predição ao quadrado}}\right]$$

MSE → Mean Squared Error

⊕ Vários algoritmos: { Regressão Linear, SVM, Árvores de regressão, Redes Neurais,

⊕ Em todos, o objetivo é encontrar uma função $g(\underset{\sim}{x})$ que minimize

$$\mathbb{E}\left[\left(Y - g(\underset{\sim}{x})\right)^2\right]$$

⊕ Minimizar num espaço de funções:

<u>Cálculo</u>

<u>Agora</u> Qual a <u>função</u> $g(\underset{\sim}{x})$ que minimiza a expressão $\mathbb{E}\left[\left(Y-g(\underset{\sim}{x})\right)^2\right]$

⊕ Temos a solução teórica e perfeita

Teorema Dentre as infinitas funções $g(\underset{\sim}{x})$, aquela que minimiza

$$\mathbb{E}\left[\left(Y - g(\underset{\sim}{x})\right)^2\right]$$

é a esperança condicional

$$\mu(\underset{\sim}{x}) = \mathbb{E}\left(Y \mid \underset{\sim}{X}\right)$$

(+) Temos a <u>solução teórica e perfeita</u>

<u>Teorema</u> Dentre as infinitas funções $g(\underset{\sim}{x})$, aquela que minimiza

$$\mathbb{E}\left[\left(Y - g(\underset{\sim}{x})\right)^2\right]$$

é a esperança condicional

$$\mu(\underset{\sim}{x}) = \mathbb{E}\left(Y \mid \underset{\sim}{X}\right)$$

Provamos este resultado na aula passada. Ver notas no website

- Esta solução teórica é pouco útil pois não sabemos calcular $\mu(\underset{\sim}{x}) = \mathbb{E}(Y|\underset{\sim}{X})$
- Em geral, não conhecemos a distribuição conjunta das variáveis aleatórias

$$(Y, \underset{\sim}{X}) = (Y, X_1, X_2, \ldots, X_{30})$$

- O que fazemos então?

—

(9)

⊕ Todo modelo/algoritmo de ML estabelece uma classe $\mathcal{C}$ (um conjunto) de funções possíveis $g(\underset{\sim}{x})$

⊕ Dentro desse ~~sub~~conjunto $\mathcal{C}$, procura achar aquela $g^*(\underset{\sim}{x})$ que seja próxima da função ótima perfeita $\mu(\underset{\sim}{x}) = \mathbb{E}(Y \mid \underset{\sim}{x})$

⊕ Exemplo :

Regressão linear

$$\mathcal{C} = \left\{ g(\underset{\sim}{x}) = \underbrace{\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \dots + \theta_{30} X_{30}}_{\text{combinações lineares das features}} \right\}$$

Cada $g(\underset{\sim}{x}) \in \mathcal{C}$ é determinada pelos coeficientes $(\theta_0, \theta_1, \dots, \theta_{30}) \in \mathbb{R}^{31}$

Queremos encontrar uma $g^*(\underset{\sim}{X}) \in \mathcal{C}$

(isto é, escolher coeficientes $\theta_0, \theta_1, \ldots, \theta_{30}$)

tal que

$$\mathbb{E}\left[\left(Y - \underbrace{(\theta_0 + \theta_1 X_1 + \ldots + \theta_{30} X_{30})}_{g(\underset{\sim}{X})}\right)^2\right]$$

seja mínimo

Queremos encontrar uma $g^*(\underset{\sim}{X}) \in \mathcal{C}$

(isto é, escolher coeficientes $\theta_0, \theta_1, \ldots, \theta_{30}$)

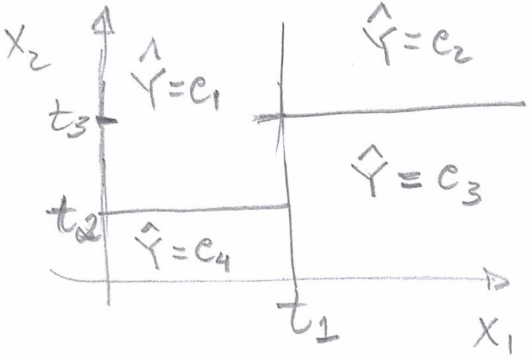
tal que

$$\mathbb{E}\left[\left(Y - \underbrace{(\theta_0 + \theta_1 X_1 + \ldots + \theta_{30} X_{30})}_{g(\underset{\sim}{X})}\right)^2\right]$$

seja mínimo

Justificativa para esta escolha linear para o conjunto C: usar a melhor aprox polinomial para E(Y | X) = mu(x). Aqui, usamos aprox do polinomio de primeira ordem (vimos na aula passada).

Exemplo: Regression tree with two features:



$\hat{Y} = g(\underset{\sim}{X}) = $ prediction

"função azulejo"

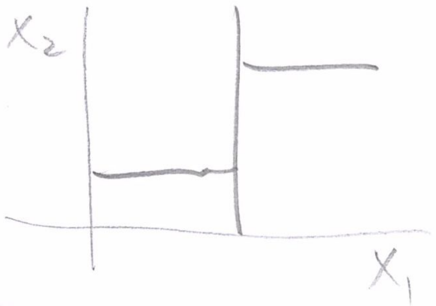
$$\hat{Y} = g(X_1, X_2) = \begin{cases} c_1; & \text{if } X_1 \leq t_1 \ \underline{\text{AND}}\ X_2 > t_2 \\ c_2, & \text{if } X_1 > t_1 \ \underline{\text{AND}}\ X_2 > t_3 \\ c_3, & \text{if } X_1 > t_1 \ \underline{\text{AND}}\ X_2 \leq t_3 \\ c_4, & \text{if } X_1 \leq t_1 \ \underline{\text{AND}}\ X_2 \leq t_2 \end{cases}$$

⊕ Achar a melhor $g(x_1, x_2)$ nesta

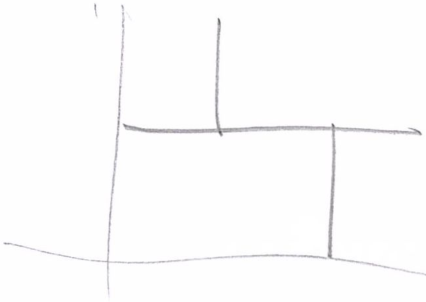
classe de "funções azulejos" $\mathcal{C}$

é encontrar os pontos de corte

t_1, t_2, t_3 e os valores c_1, c_2, c_3, c_4



ou



ou ...

RESUMO Objetivo: achar a funç. de $\underset{\sim}{X}$ ótima (para minimizar o (erro)^2 esperado)

$$\mu(\underset{\sim}{x}) = \mathbb{E}(Y \mid \underset{\sim}{X})$$

Soluç. Definir uma classe de funções $\mathcal{C}$ que seja:

trade-off
- ⊕ flexível, maleável, que possa aproximar qualquer funç.
- ⊖ "pequena" para ser trabalhada numericamente

• Soluções: SVM, NN, Regressão linear...

(15)

Na classe $\mathcal{G}$ escolhida pelo algoritmo, achar a função $g^*(x) \in \mathcal{G}$ que melhor aproxime $\mu(x) = \mathbb{E}(Y|\underset{\sim}{x})$

Mas como encontrar esta $g^*(\underset{\sim}{x})$ se não conhecemos $\mu(\underset{\sim}{x})$?

Na prática, buscamos minimizar a soma dos erros de predição no conjunto de treinamento

Para qualquer vetor aleatório

$$(Y, \underset{\sim}{X}) = (Y, X_1, \ldots, X_{30})$$

o valor esperado de uma função $h(Y, \underset{\sim}{X})$ é aprox. a média aritmética de h calculada numa amostra de tamanho grande:

$$\mathbb{E}\left[h(Y, \underset{\sim}{X})\right] \approx \frac{1}{N} \sum_{i=1}^{N} h(Y_i, \underset{\sim}{X}_i)$$

$$\mathbb{E}\Big(\underbrace{\big(Y - g(\underset{\sim}{x})\big)^2}_{h(Y, \underset{\sim}{x})} \Big) \approx \underbrace{\frac{1}{N}\sum_{i=1}^{N}\big(Y_i - g(\underset{\sim}{x}_i)\big)^2}_{\text{Ache } g^* \in \mathcal{G} \text{ que minimiza esta soma}}$$

- Como minimizar um objetivo num espaço de funções? Derivar e igualar a zero? Derivar em relação a quê?

Se $\mathcal{C}$ é uma classe parametrizada por um vetor de coeficientes $(\theta_0, \theta_1, \ldots, \theta_{30})$, derive em relação a estes coeficientes:

Ex: Regressão Linear

$$g(\underset{\sim}{x}) = \beta_0 + \beta_1 x$$

Achar $\boxed{\beta_0 \quad e \quad \beta_1}$ que minimizam

$$\frac{1}{N}\sum_{i=1}^{N}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 = \frac{1}{N}\left((70.3 - (\beta_0 + \beta_1 150))^2 + (90.7 - (\beta_0 + \beta_1 210))^2 + \ldots \right)$$

Ignorando a constante 1/N, derive em relação a beta_0 e beta_1 e iguale a zero:

- Derivando e igualando a zero:

$$
\begin{aligned}
0 &= \sum_i \frac{\partial}{\partial \beta_0} (y_i - (\beta_0 + \beta_1 x_i))^2 \\
0 &= \sum_i \frac{\partial}{\partial \beta_1} (y_i - (\beta_0 + \beta_1 x_i))^2
\end{aligned}
$$

- Ou seja:

$$
\begin{aligned}
0 &= \sum_i 2\,(y_i - (\beta_0 + \beta_1 x_i))\ (-1) \\
0 &= \sum_i 2\,(y_i - (\beta_0 + \beta_1 x_i))\ (-x_i)
\end{aligned}
$$

- Temos

$$\begin{aligned}
0 &= -\sum_i y_i + \beta_0\ n + \beta_1 \sum_i x_i \\
0 &= -\sum_i (y_i x_i) + \beta_0\ \sum_i x_i + \beta_1 \sum_i x_i^2
\end{aligned}$$

- Rearranjando:

$$\begin{aligned}
\beta_0\ n + \beta_1\ \sum_i x_i &= \sum_i y_i \\
\beta_0\ \sum_i x_i + \beta_1\ \sum_i x_i^2 &= \sum_i (y_i x_i)
\end{aligned}$$

- Este é um sistema linear de duas equações com duas incógnitas, β_0 e β_1.

- Sistema na forma matricial

$$\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i (x_i y_i) \end{bmatrix}$$

- Com solução:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i y_i \\ \sum_i (x_i y_i) \end{bmatrix}$$

- Vamos usar uma notação para simplificar as expressões.
- Vamos denotar a média dos x e y's por
  - $\bar{x} = \dfrac{1}{n}\sum_i x_i$, média aritmética dos x_i's
  - $\bar{y} = \dfrac{1}{n}\sum_i y_i$
  - $\overline{x^2} = \dfrac{1}{n}\sum_i x_i^2$, média aritmética dos x_i^2's
  - $\overline{xy} = \dfrac{1}{n}\sum_i (x_i y_i)$, média aritmética dos $x_i y_i$'s

- Sistema na forma matricial e com a notação introduzida

$$\begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix}$$

- Com solução:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}^{-1} \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix}$$

- Como a inversa de uma matriz 2×2 é conhecida, podemos resolver de forma explícita a solução de mínimos quadrados.
- Após alguma manipulação algébrica, temos a solução como uma fórmula envolvendo os pontos:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \, \bar{x}
\end{aligned}
$$

Regressão: caso geral

- Vetor Y de dimensão n:
  - preço de n aptos
- (k+1) features, vetores-coluna de dimensão n
  - Feature 0: vetor-coluna de 1's
  - Feature 1: vetor-coluna com área dos aptos n aptos
  - ….
  - Feature k: vetor-coluna com indicador binário "tem salão de festa?"
- Colete as features numa matriz X de dimensão $n \times (k+1)$
- Objetivo: minimizar

$$\sum_{i=1}^{n}(y_i - \beta_0 * 1 + \beta_1 x_{i1} + \ldots \beta_k x_{ik})^2 = \sum_{i=1}^{n} ||\mathbf{Y} - \mathbf{X}\beta||^2$$

Regressão: caso geral

- Derive
- $\sum_{i=1}^{n}(y_i - \beta_0 * 1 + \beta_1 x_{i1} + \ldots \beta_k x_{ik})^2 = \sum_{i=1}^{n} ||\mathbf{Y} - \mathbf{X}\beta||^2$
- em relação aos coeficientes β_j de cada feature
- Iguale cada derivada parcial a zero
- “Isole” os beta’s
- Teremos um sistema linear chamado equações normais:

$$(\mathbf{X}^t\mathbf{X})\beta = \mathbf{X}^t\mathbf{Y}$$

Equações normais

$$\underbrace{(\mathbf{X}^t\mathbf{X})}_{(k+1)\times n \quad n\times(k+1)} \quad \underbrace{\beta}_{(k+1)\times 1} = \underbrace{\mathbf{X}^t\mathbf{Y}}_{(k+1)\times n \quad n\times 1}$$

$$\underbrace{(\mathbf{X}^t\mathbf{X})}_{(k+1)\times(k+1)} \quad \underbrace{\beta}_{(k+1)\times 1} = \underbrace{\mathbf{X}^t\mathbf{Y}}_{(k+1)\times 1}$$

Temos um sistema linear

X'X é matriz quadrada (k+1) x (k+1)

Se tiver inversa, podemos obter a solução

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

Uma abordagem mais geral

Solução Least Squares foi vista como um problema numérico:

minimizar uma função objetivo

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \sum_i \left(\text{preço}_i - \beta_0 * 1 + \beta_1 \text{área}_i + \ldots + \beta_k \text{banheiro?}_i\right)^2$$

Ao invés de olhar o problema como um problema numérico, vamos dar uma roupagem mais teórica.

Vai permitir uma solução mais elegante e … generalizável para espaços de funções

- Mas, e quando $\mathcal{C}$ não é tão simples?
- O que é uma classe $\mathcal{C}$ boa?
- Podemos garantir que <u>sempre</u> vamos encontrar uma função $g^*(\underset{\sim}{x})$ que seja uma boa aproximação para $\mu(\underset{\sim}{x}) = \mathbb{E}(Y|\underset{\sim}{X})$?
- Precisamos de três coisas:
  - trabalhar no espaço de funções
  - medir distância entre funções
  - saber quando uma sequência de funções converge (aproxima) para outra

- Isto nos leva a conceitos de <u>análise funcional</u>
- Recordar <u>espaço vetorial</u>

Conjunto de objetos em que:

- podemos <u>somar objetos</u> e obter um <u>3º objeto</u>
- podemos "espichar", "encolher" ou "reverter" objetos ao multiplicar <u>por escalares</u>

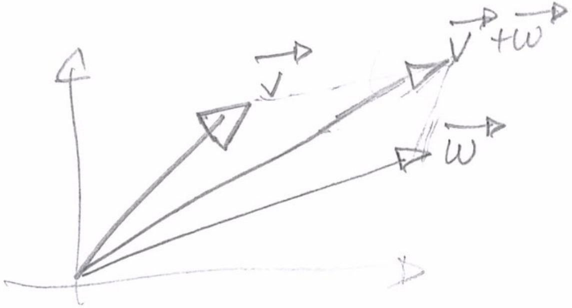
To have a vector space, the eight following axioms must be satisfied for every $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ in V, and a and b in F.[3]

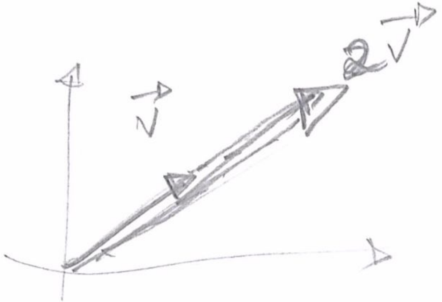
Axiom	Statement
Associativity of vector addition	$\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
Commutativity of vector addition	$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
Identity element of vector addition	There exists an element $\mathbf{0} \in V$, called the *zero vector*, such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for all $\mathbf{v} \in V$.
Inverse elements of vector addition	For every $\mathbf{v} \in V$, there exists an element $-\mathbf{v} \in V$, called the *additive inverse* of $\mathbf{v}$, such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$.
Compatibility of scalar multiplication with field multiplication	$a(b\mathbf{v}) = (ab)\mathbf{v}$ [nb 3]
Identity element of scalar multiplication	$1\mathbf{v} = \mathbf{v}$, where 1 denotes the multiplicative identity in F.
Distributivity of scalar multiplication with respect to vector addition	$a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$
Distributivity of scalar multiplication with respect to field addition	$(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$

Exemplos de espaços vetoriais:

$$\vec{v} \in \mathbb{R}^2 = \{ \vec{v} = (x, y), \quad x \in \mathbb{R}, \ y \in \mathbb{R} \}$$

com escalares reais

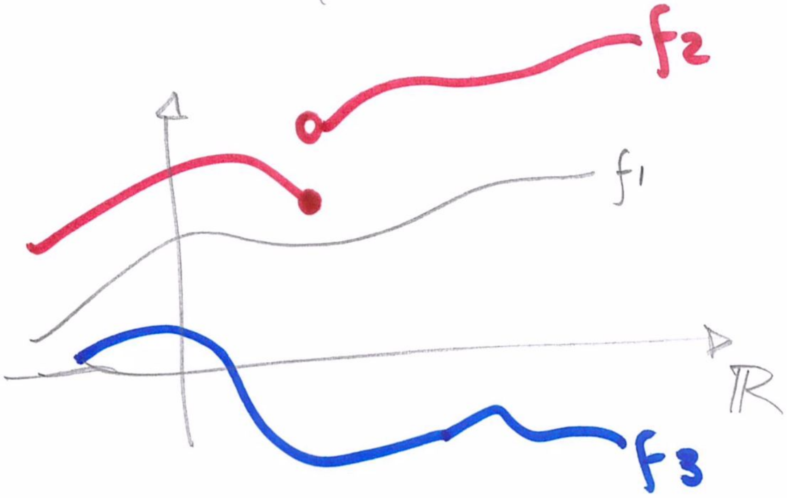




- $\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n),\ x_i \in \mathbb{R}\}$

- <u>Espaço de funções</u> $\mathcal{F}$

$$\mathcal{F} = \{ f : \mathbb{R} \to \mathbb{R} \}$$

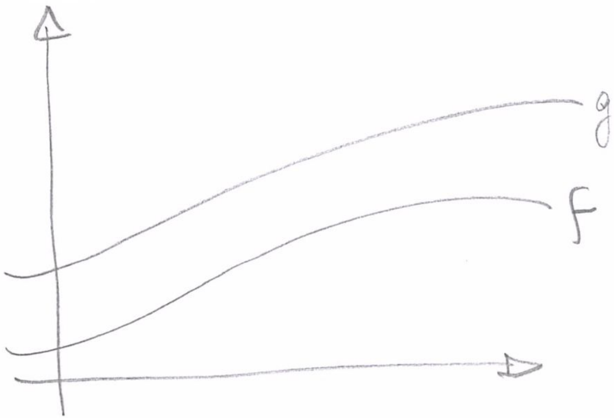


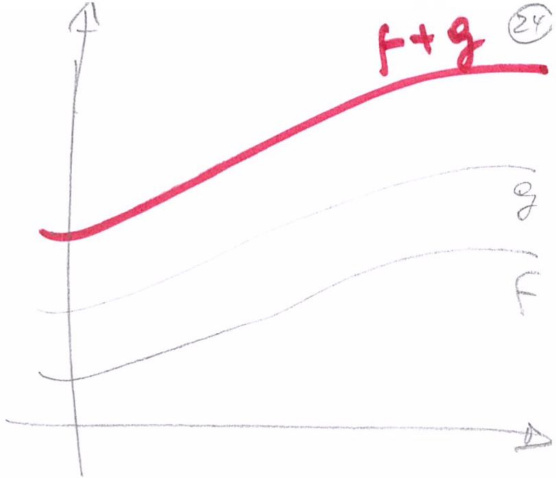
somar duas funções $\rightarrow$ ainda é uma função 23

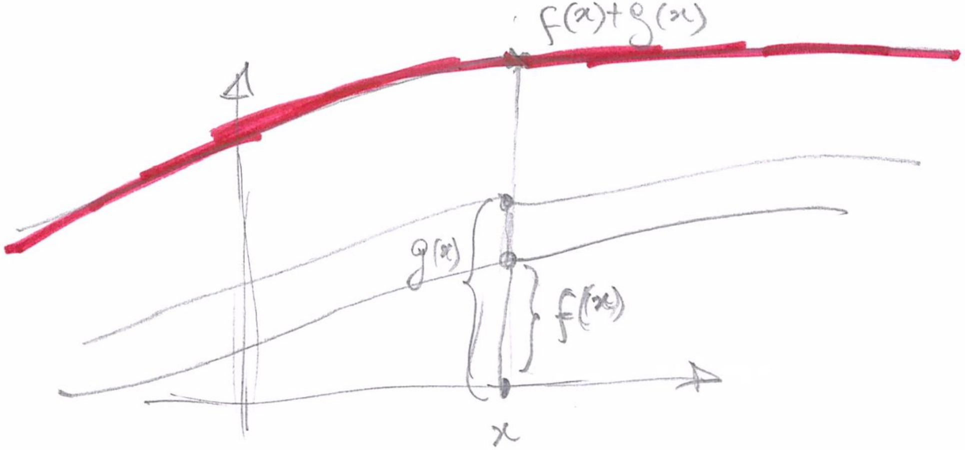
Definir a soma $g+f = h$:

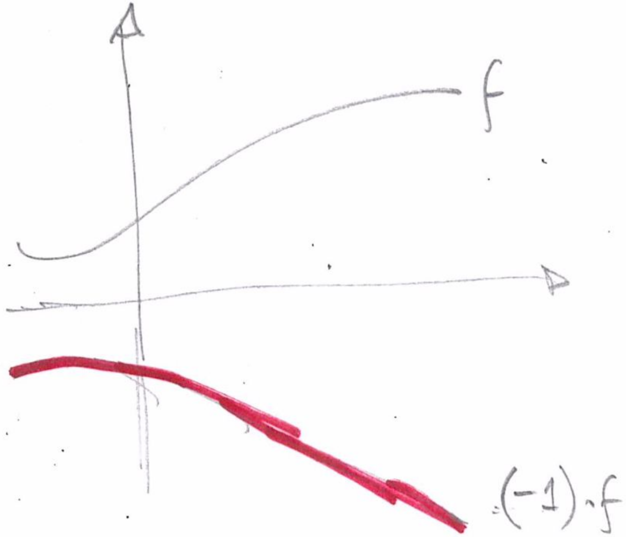
é a função tal que

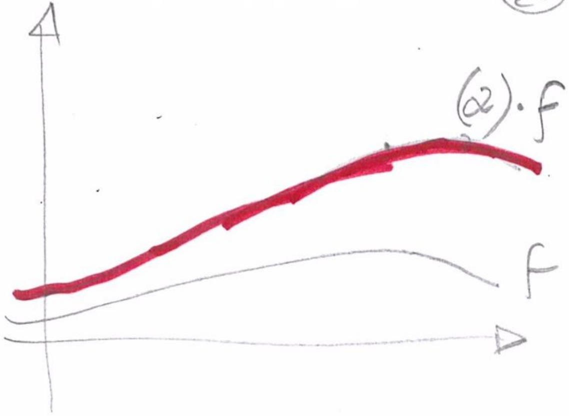
$h(x) = g(x) + f(x)$ para todo x

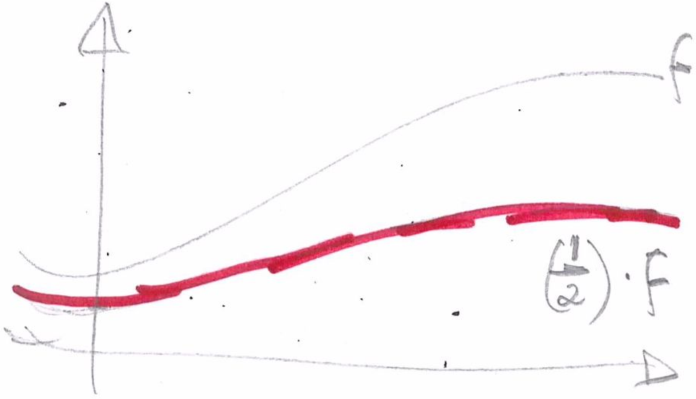

g
f


f+g
g
f


f(x)+g(x)
g(x)
f(x)
x

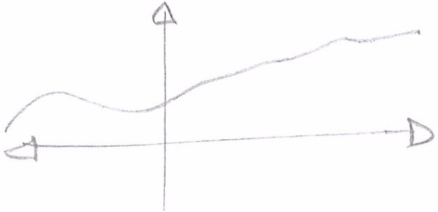

f
(-1)·f


(2)·f
f

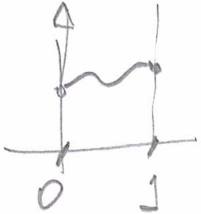

f
(1/2)·f

$\mathcal{F} = \{ f$ funções em qual domínio? $\}$

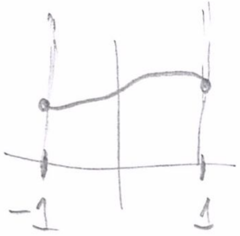
$\mathcal{F}(\mathbb{R}) = \{ f : \mathbb{R} \longrightarrow \mathbb{R} \}$



$\mathcal{F}([0,1]) = \{ f : [0,1] \longrightarrow \mathbb{R} \}$

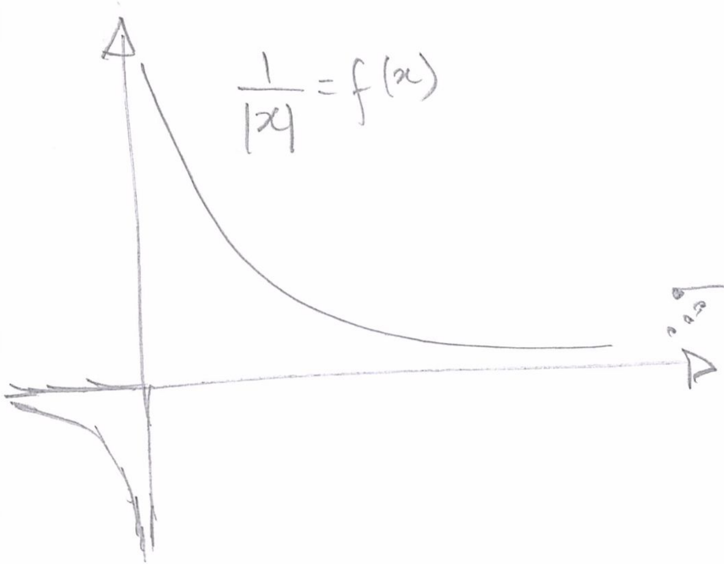


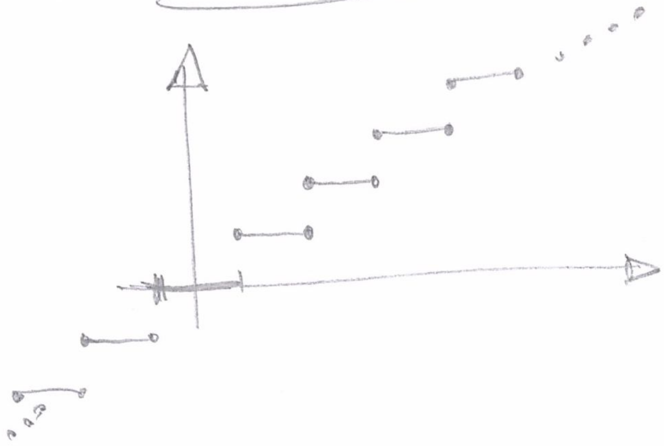
$\mathcal{F}([-1,1]) = \{ f : [-1,1] \longrightarrow \mathbb{R} \}$



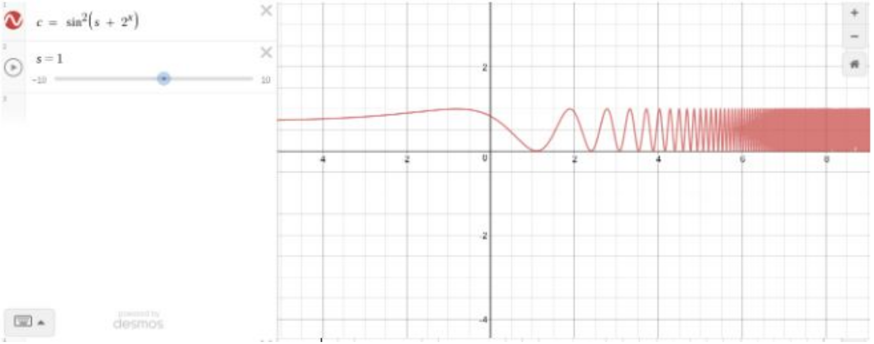
$$\mathcal{F}(I) = \{ f : \text{Intervalo } I \longrightarrow \mathbb{R} \}$$

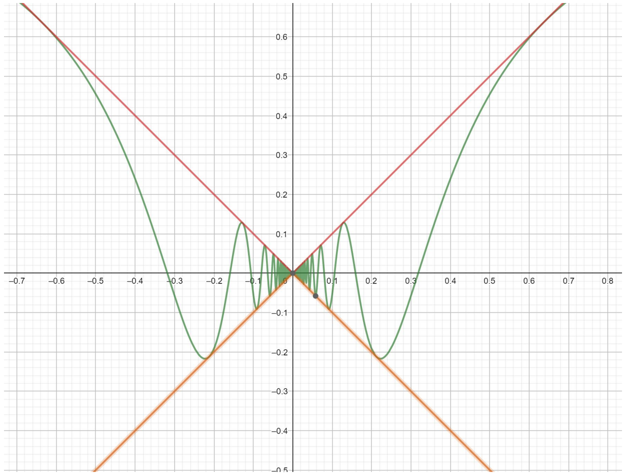
$\mathcal{F}(I)$ tem elementos <u>muito estranhos</u>

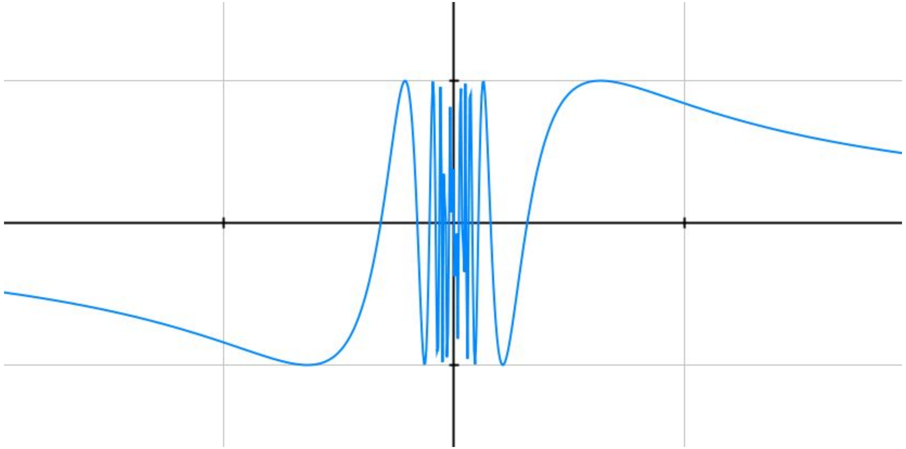


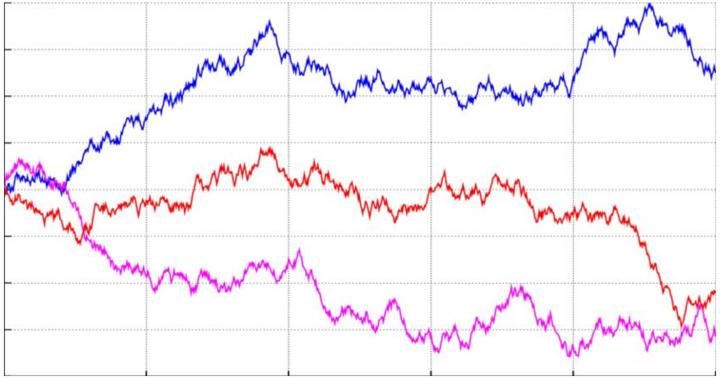


$c = \sin^2(s + 2^x) \qquad s = 1$









Ignorar espaços de funções (por enquanto)
Foco em $\mathbb{R}^n$ e regressão linear

Exemplo de preço de apto

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

- Y é um vetor de dimensão 1500 escrito como combinação linear de 31 vetores, cada um deles de dimensão 1500.
- Problema: encontrar os coeficientes $b_0, b_1, \ldots, b_{30}$ que tornem a aproximação acima a melhor possível.

A matriz de desenho X

- Seja X a matriz 1500×31 abaixo (note que ela tem uma coluna composta apenas de 1's):

$$X = \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix}$$

Combinações lineares e a matriz X

- A combinação linear que buscamos

$$b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

- pode ser escrita como

$$X\ b = \begin{bmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_30 \end{bmatrix}$$

Vetores próximos

Nosso problema é encontrar os coeficientes $b_0, b_1, \ldots b_{30}$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Ou seja, encontrar $b_0, b_1, \ldots b_{30}$ tais que $Y \approx Xb$ onde

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{1498} \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{30} \end{pmatrix} = Xb$$

onde $b = (b_0, \ldots b_{30})^t$.

Solução: minimizar norma

- X é uma matriz 1500×31.
- Y e Xb são vetores 1500-dim.
- Além disso, Xb é uma combinação linear das colunas da matriz X.
- Queremos encontrar b tal que o vetor Xb seja o mais próximo possível do vetor Y.
- Queremos $Y - Xb$ aproximadamente igual AO VETOR ZERO.
- Queremos $||Y - Xb|| \approx 0$ (o comprimento-norma é um número, não um vetor)

Solução melhor: minimizar norma ao quadrado

- Queremos $||Y - Xb|| \approx 0$
- Queremos $\widehat{b}$ que minimize $||Y - Xb||$
- Mas norma euclidiana envolve a raiz quadrada da soma dos quadrados ...
- Mas se $\widehat{b}$ minimiza $||Y - Xb||$ então $\widehat{b}$ minimiza $||Y - Xb||^2$
- Esta segunda função é mais fácil de derivar.

Solução melhor: minimizar norma ao quadrado

- Então procuramos vetor b tal que $||Y - Xb||^2 \approx 0$.
- Queremos $\widehat{b}$ que minimize $||Y - Xb||^2$
- Matematicamente: queremos $\widehat{b} = \arg\min_{b} ||Y - Xb||^2$.
- Como encontrar este $\widehat{b}$?

Vetores e combinações lineares

- X é matriz 1500×31. b é vetor 31×1
- Para qualquer vetor $b \in \mathbb{R}^{31}$, temos Xb em $\mathbb{R}^{1500}$.
- Varie b varrendo todos os vetores b possíveis. O que obtemos?
- Isto é, o que é o conjunto

$$\mathfrak{M}(X) = \{v \in \mathbb{R}^{1500} \text{ tais que } v = Xb \text{ para algum } b\} \quad ?$$

O que é $\mathfrak{M}(X)$?

- Colunas da matriz X estão fixadas, são vetores 1500×1 de números constantes, conhecidos.

$$\mathfrak{M}(X) = \left\{ b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix} \right\}$$

- $\mathfrak{M}(X)$ é um subconjunto de vetores do espaço vetorial $\mathbb{R}^{1500}$.
- Vetor zero pertence a $\mathfrak{M}(X)$.
- Somando duas combinações lineares de $\mathfrak{M}(X)$ ainda permanecemos em $\mathfrak{M}(X)$.
- Multiplicando um elemento de $\mathfrak{M}(X)$ por uma constante ainda permanecemos em $\mathfrak{M}(X)$.

Sub-espaços vetoriais

- Informal: Sub-espaço vetorial W de um espaço vetorial V é um subconjunto de V tal que:
  - a soma de dois vetores de W permanece em W
  - multiplicar um vetor de W por um escalar permanece em W
  - O vetor 0 (nulo) pertence a W

- Só isto: W é um espaço vetorial e não saímos de dentro dele ao manusear seus vetores com adição ou multiplicação por escalar.

Espaço $\mathfrak{M}(X)$ das combinações lineares

- $\mathfrak{M}(X)$ é um sub-espaço vetorial de $\mathbb{R}^{1500}$.
- $\mathfrak{M}(X)$ é o sub-espaço vetorial formado pelas combinações lineares dos 31 vetores-colunas de X.
- Se as colunas de X são linearmente independentes, então $\mathfrak{M}(X)$ é um sub-espaço vetorial de dimensão igual ao número de colunas de X (que é 31, no nosso exemplo).
- Nosso problema então é: encontrar os coeficientes b da combinação linear $Xb \in \mathfrak{M}(X)$ tal que Xb seja o mais próximo posível do vetor Y.

Geometria dos Mínimos quadrados

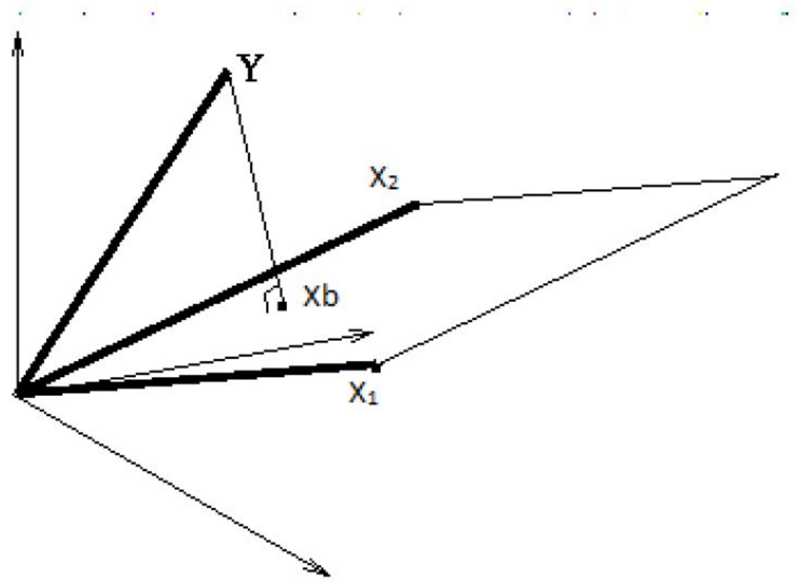


Figura: Representação do vetor $Y \in \mathbb{R}^{1500}$. O plano inclinado representa o sub-espaço vetorial $\mathfrak{M}(X)$ gerado por uma matriz X com apenas duas colunas, os vetores X_1 e X_2, ambos do $\mathbb{R}^{1500}$. O sub-espaço vetorial $\mathfrak{M}(X)$ é de dimensão 2. Identifique visualmente o ponto-vetor em $\mathfrak{M}(X)$ que minimiza $||Y - Xb||^2$.

Teorema da Projeção Ortogonal

- Seja $\mathbb{R}^n$ um espaço vetorial real de dimensão n.
- Seja $\mathcal{W}$ um sub-espaço vetorial de $\mathcal{V}$ com dimensão m.
- Seja $Y \in \mathbb{R}^n$ um vetor qualquer.
- **Teorema:** Existe um único vetor $\hat{w} \in \mathcal{W}$ que minimiza $||Y - w||$ com $w \in \mathcal{W}$.
- Além disso, este $\hat{w} \in \mathcal{W}$ é o único vetor tal que $Y - \hat{w}$ é ortogonal a $\hat{w}$. Isto é, $\hat{w}$ é o único vetor tal que $(Y - \hat{w}) \perp \hat{w}$.

PROVA: a seguir

Teorema da Projeção Ortogonal

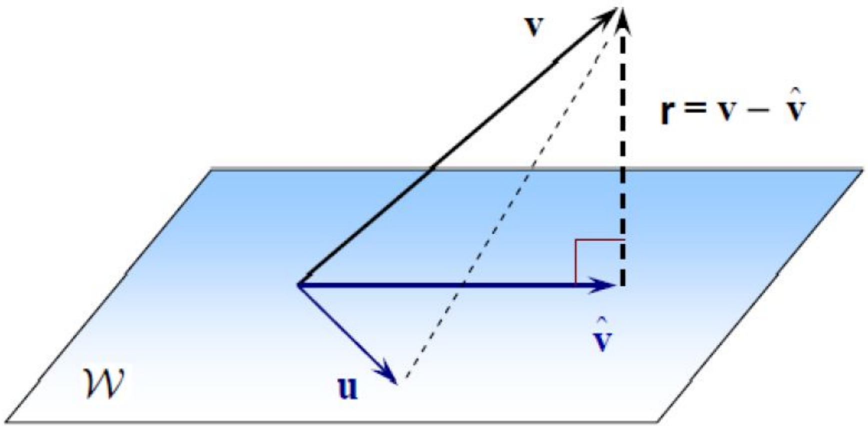


Figura: **v** é um vetor do $\mathbb{R}^3$. O plano $\mathcal{W}$ é um sub-espaço vetorial de dimensão 2. dado um vetor **u** do sub-espaço, $||\mathbf{v} - \mathbf{u}||$ (linha tracejada fina) é o comprimento do vetor $\mathbf{v} - \mathbf{u}$.

De todos os vetores **u** do sub-espaço $\mathcal{W}$, aquele que minimiza o comprimento $||\mathbf{v} - \mathbf{u}||$ é a projeção ortogonal $\hat{\mathbf{v}}$. O vetor $\hat{\mathbf{v}}$ é a aproximação de mínimos quadrados em $\mathcal{W}$ para **v**. O vetor $r = \mathbf{v} - \hat{\mathbf{v}}$ é o vetor de resíduos.

Geometria dos Mínimos quadrados

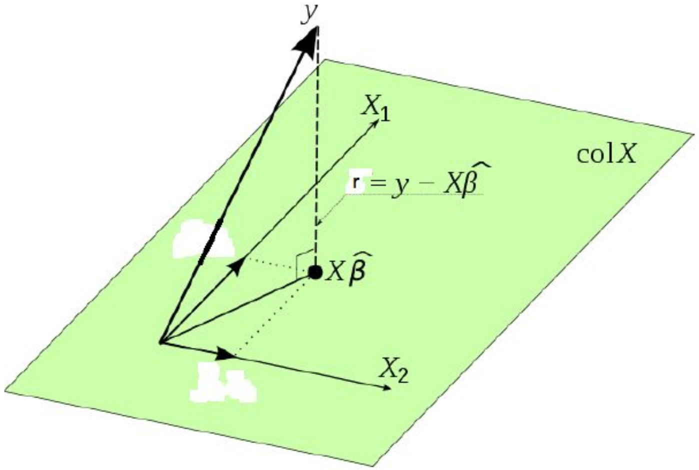


Figura: Projeção ortogonal de $Y \in \mathbb{R}^{1500}$ no sub-espaço vetorial $\mathfrak{M}(X)$ minimiza $||Y - X\beta||^2$. Esta projeção é o vetor $X\hat{\beta}$. Vetor de resíduos é $r = Y - X\hat{\beta}$ e é $\perp$ a $X\hat{\beta}$. Imagem retirada de
`https://commons.wikimedia.org/w/index.php?curid=7309159`

- Demonstração do Teorema da Projeção.
- Não veremos a demonstração geral para espaços vetoriais arbitrários.
- Vamos fazer apenas o caso especial da regressão linear.

Produto interno com vetores-coluna

- Produto interno de dois vetores, $\mathbf{v}$ e $\mathbf{w}$, no mesmo espaço vetorial de dimensão n:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{n} v_i w_i$$

- Como $\mathbf{v}$ e $\mathbf{w}$ são vetores-coluna, temos

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i} v_i w_i = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \mathbf{v}^t \mathbf{w}$$

- Temos $\mathbf{v} \perp \mathbf{w}$ se, e só se, $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^t\ \mathbf{w} = 0$.

- Imitando o Scilab, vamos denotar $\mathbf{v}^t$ por $\mathbf{v}'$.

- Primeiro, queremos achar $\hat{Y} = X\hat{\beta}$ tal que $\left(Y - \hat{Y}\right) \perp \hat{Y}$
- Depois, queremos mostrar que este $\hat{Y} = X\hat{\beta}$ é o vetor que minimiza a distância $||Y - X\beta||^2$

$$
\begin{aligned}
\hat{Y} &= X\beta \perp \left(Y - \hat{Y}\right), \ \forall Y \\
\left\langle X\beta, Y - \hat{Y} \right\rangle &= 0 \\
0 &= (X\beta)^t (Y - X\beta) \\
&= \beta^t X^t (Y - X\beta) \\
&= \beta^t (X^t Y - X^t X\beta)
\end{aligned}
$$

- Ou $\beta^t = 0$ (o que implica que $\beta = 0$ e que $\hat{Y} = 0$), solução sem sentido
- Ou $X^t Y - X^t X\beta = 0 \Longrightarrow$ com solução $\hat{\beta} = \left(X^t X\right)^{-1} X^t Y$

- Seja $\hat{Y} = X\hat{\beta} = X\left(X^tX\right)^{-1}X^tY$
- Vamos calcular agora $||Y - X\beta||^2$ para um β arbitrário:
- Some e subtraia: $Y - X\beta = Y - X\hat{\beta} + X\hat{\beta} - X\beta$
- Calculando:

$$
\begin{aligned}
&||Y - X\beta||^2 = (Y - X\beta)^t(Y - X\beta) = \\
&= \left(\left(Y - X\hat{\beta}\right) + \left(X\hat{\beta} - X\beta\right)\right)^t\left(\left(Y - X\hat{\beta}\right) + \left(X\hat{\beta} - X\beta\right)\right) \\
&= \left(Y - X\hat{\beta}\right)^t\left(Y - X\hat{\beta}\right) + \left(Y - X\hat{\beta}\right)^t\left(X\hat{\beta} - X\beta\right) + \\
&+ \left(X\hat{\beta} - X\beta\right)^t\left(Y - X\hat{\beta}\right) + \left(X\hat{\beta} - X\beta\right)^t\left(X\hat{\beta} - X\beta\right) \\
&= \left\|Y - X\hat{\beta}\right\|^2 + \underbrace{2\left(Y - X\hat{\beta}\right)^t\left(X\hat{\beta} - X\beta\right)}_{A} + \left\|X\hat{\beta} - X\beta\right\|^2
\end{aligned}
$$

- Vamos mostrar agora que A $=0$

$$\left(Y - X\hat{\beta}\right)^t\left(X\hat{\beta} - X\beta\right) = \left(Y - X\left(X^tX\right)^{-1}X^tY\right)^t\left(X\hat{b} - Xb\right)$$
$$= \left(\left(I - X\left(X^tX\right)^{-1}X^t\right)Y\right)^t\left(X\hat{b} - Xb\right)$$
$$= \left(Y^t\left(I - X\left(X^tX\right)^{-1}X^t\right)^t\right)\left(X\left(\hat{\beta} - \beta\right)\right) = (*)$$

- Temos $\left(I - X\left(X^tX\right)^{-1}X^t\right)^t = I^t - \left(X^t\right)^t\left(\left(X^tX\right)^{-1}\right)^tX^t$

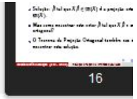
$$= \quad I - X\left(\left(X^tX\right)^t\right)^{-1}X^t$$
$$= \quad I - X\left(X^tX\right)^{-1}X^t$$

- Assim $(*) = Y^t\left(I - X\left(X^tX\right)^{-1}X^t\right)X\left(\hat{\beta} - \beta\right)$

$$
\begin{aligned}
&= Y^t\left[\left(I - X\left(X^tX\right)^{-1}X^t\right)X\right]\left(\hat{\beta} - \beta\right) \\
&= Y^t\left[X - X\underbrace{\left(X^tX\right)^{-1}X^tX}_{I}\right]\left(\hat{\beta} - \beta\right) \\
&= Y^t\left[X - X\right]\left(\hat{\beta} - \beta\right) \\
&= Y^t\ \ [0]\ \ \left(\hat{\beta} - \beta\right) = 0
\end{aligned}
$$

- Portanto:
$||Y - X\beta||^2 = \left|\left|Y - X\hat{\beta}\right|\right|^2 + 0 + \left|\left|X\hat{\beta} - X\beta\right|\right|^2 \geq \left|\left|Y - X\hat{\beta}\right|\right|^2$
- pois, sendo uma distância, $\left|\left|X\hat{\beta} - X\beta\right|\right|^2$ é sempre maior ou igual a zero.
- Isto é, $\hat{\beta} = (X^tX)^{-1}X^tY$ é o vetor de coeficientes β que minimiza $||Y - X\beta||^2$ para todo vetor Y

A solução de mínimos quadrados



- Nosso problema: encontrar $\hat{\beta}$ tal que o vetor $X\beta$ do subespaço $\mathfrak{M}(X)$ seja o mais próximo posível do vetor Y.
- O Teorema da Projeção Ortogonal garante que existe uma solução. Além disso,...
- Solução: $\hat{\beta}$ tal que $X\hat{\beta} \in \mathfrak{M}(X)$ é a projeção ortogonal de Y em $\mathfrak{M}(X)$.
- Mas como encontrar este vetor $\hat{\beta}$ tal que $X\hat{\beta}$ e seja esta projeção ortogonal?
- O Teorema da Projeção Ortogonal também nos dá a dica de como encontrar esta solução.

Encontrando a solução de mínimos quadrados

- Solução é $\hat{\beta}$, um vetor tal que $X\hat{\beta} \in \mathfrak{M}(X)$ é a projeção ortogonal de Y em $\mathfrak{M}(X)$.
- O Teorema da Projeção Ortogonal diz que a projeção $X\hat{\beta}$ é única e é o vetor tal que $X\hat{\beta} \perp (Y - X\hat{\beta})$.
- Em resumo, devemos ter o produto interno zerado: $\langle X\hat{\beta}, (Y - X\hat{\beta}) \rangle = 0$.
- Isto implica que

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Example

Concrete Compressive Strength

- A Resistência à Compressão do Concreto é uma medida da capacidade do concreto de suportar cargas que tendem a comprimir ou reduzir seu volume.
- É uma das propriedades mais críticas do concreto, indicando sua capacidade de suportar cargas estruturais sem romper.
- É definida como a tensão compressiva máxima que o concreto pode resistir antes da falha.
- É determinada pela aplicação de uma carga compressiva a uma amostra de concreto (geralmente um cubo ou um cilindro) sob condições controladas até que ela se quebre.
- https://www.youtube.com/shorts/yGpwrL0zwHI

Kaggle Dataset

- Aim: To predict the compressive strength of concrete based on material composition.

🎯 Target Variable (Response Variable)

Feature Name	Description	Units	Typical Range
Compressive Strength	The maximum compressive stress the concrete can withstand.	MPa (MegaPascals)	2.33 - 82.6

- Number of Samples: 1,030 observations
- Number of Features: 8 predictors

🔍 Features (Predictor Variables)

Feature Name	Description	Units	Typical Range
Cement	The amount of cement used in the mix.	kg/m^3	102 - 540
Blast Furnace Slag	By-product of steel production, often used as a cement substitute.	kg/m^3	0 - 359.4
Fly Ash	A by-product of coal combustion, used as a partial cement replacement.	kg/m^3	0 - 200.1
Water	The amount of water used in the mix.	kg/m^3	121.8 - 247
Superplasticizer	Chemical additive to enhance workability and strength.	kg/m^3	0 - 32.2
Coarse Aggregate	Gravel or crushed stone used as a filler material.	kg/m^3	801 - 1145
Fine Aggregate	Sand used as a filler material.	kg/m^3	594 - 992.6
Age	Age of the concrete sample when tested.	days	1 - 365

```
X'X matrix (9x9) scaled by 10^7:
 [[ 0.    0.03  0.01  0.01  0.02  0.    0.1   0.08  0.  ]
 [ 0.03  9.27  1.88  1.3   5.24  0.19 28.08 22.21  1.38]
 [ 0.01  1.88  1.33  0.23  1.4   0.05  7.21  5.69  0.32]
 [ 0.01  1.3   0.23  0.72  0.98  0.05  5.43  4.36  0.19]
 [ 0.02  5.24  1.4   0.98  3.44  0.11 18.16 14.39  0.89]
 [ 0.    0.19  0.05  0.05  0.11  0.01  0.61  0.51  0.02]
 [ 0.1  28.08  7.21  5.43 18.16  0.61 98.12 77.41  4.57]
 [ 0.08 22.21  5.69  4.36 14.39  0.51 77.41 62.3   3.56]
 [ 0.    1.38  0.32  0.19  0.89  0.02  4.57  3.56  0.63]]
X'Y vector (9x1) scaled by 10^7:  [0.   1.13 0.29 0.19 0.66 0.03 3.57 2.83 0.2 ]

The Normal Equations are: X'X * B = X'Y
Where B is the vector of regression coefficients (intercept + slopes).
```

```python
#generate OLS regression results for all features
import statsmodels.api as sm

X_sm = sm.add_constant(X)
model = sm.OLS(y,X_sm)
print(model.fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  csMPa   R-squared:                       0.616
Model:                            OLS   Adj. R-squared:                  0.613
Method:                 Least Squares   F-statistic:                     204.3
Date:                Fri, 15 Oct 2021   Prob (F-statistic):          6.29e-206
Time:                        16:43:15   Log-Likelihood:                -3869.0
No. Observations:                1030   AIC:                             7756.
Df Residuals:                    1021   BIC:                             7800.
Df Model:                           8
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              -23.3312     26.586     -0.878      0.380     -75.500      28.837
cement               0.1198      0.008     14.113      0.000       0.103       0.136
slag                 0.1039      0.010     10.247      0.000       0.084       0.124
flyash               0.0879      0.013      6.988      0.000       0.063       0.113
water               -0.1499      0.040     -3.731      0.000      -0.229      -0.071
superplasticizer     0.2922      0.093      3.128      0.002       0.109       0.476
coarseaggregate      0.0181      0.009      1.926      0.054      -0.000       0.037
fineaggregate        0.0202      0.011      1.887      0.059      -0.001       0.041
age                  0.1142      0.005     21.046      0.000       0.104       0.125
------------------------------------------------------------------------------
```

Projeção e predição

- Se as variáveis (colunas) da matriz X realmente servirem para predizer o valor de Y e
- se o modelo de regressão linear for uma boa aproximação para o relacionamento das variáveis,
- então esperamos que
$$Y \approx \hat{Y} = X\hat{\beta}$$
- Como medir o grau de aproximação?
- É possível obter uma decomposição do vetor Y em componentes ortogonais. A partir daí extraímos uma medida de qualidade do ajuste.

Decomposição em soma de quadrados

- Seja $\bar{y} = \sum_i y_i/1500$, o preço médio dos 1500 apartamentos.
- Defina o vetor 1500×1 dado por $\bar{Y} = (\bar{y}, \bar{y}, \ldots, \bar{y})' = \bar{y}(1, 1, \ldots, 1)'$
- O vetor Y pode ser escrito como

$$Y = \hat{Y} + (Y - \hat{Y}) = \bar{Y} + \hat{Y} - \bar{Y} + (Y - \hat{Y})$$

- Isto é,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \end{bmatrix} + \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{1499} \\ \hat{y}_{1500} \end{bmatrix} - \begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \\ \bar{y} \end{bmatrix} + \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_{1499} - \hat{y}_{1499} \\ y_{1500} - \hat{y}_{1500} \end{bmatrix}$$

Decomposição em soma de quadrados

- Isto é,

$$\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_{1499} - \bar{y} \\ y_{1500} - \bar{y} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \vdots \\ \hat{y}_{1499} - \bar{y} \\ \hat{y}_{1500} - \bar{y} \end{bmatrix} + \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_{1499} - \hat{y}_{1499} \\ y_{1500} - \hat{y}_{1500} \end{bmatrix}$$

- $\mathbf{Y} - \bar{y}\mathbf{1} = (\hat{\mathbf{Y}} - \bar{y}\mathbf{1}) + (\mathbf{Y} - \hat{\mathbf{Y}})$
- Os vetores do lado direto são ortogonais um ao outro. Em consequência,

$$||\mathbf{Y} - \bar{y}\mathbf{1}||^2 = ||\hat{\mathbf{Y}} - \bar{y}\mathbf{1}||^2 + ||\mathbf{Y} - \hat{\mathbf{Y}}||^2$$

(a)

Relembrar

$$\underset{(p+1)\times 1}{\hat{\beta}} = (X'X)^{-1}X'Y$$

$$\underset{n\times 1}{\hat{Y}} = X\hat{\beta} = \underbrace{X\,(X'X)^{-1}X'}_{H}Y = HY$$

H = hat matrix (put the hat on top of Y)

H is $n \times n$ ~~Projection matrix~~

~~Properties of H:~~

Properties of $H = X(X'X)^{-1}X'$

i) H is $n \times n$

ii) H is the $\perp$ projection matrix

$\perp$ Projects where? $\longrightarrow$ Im $\mathcal{M}(X)$

linear combinatio[n]
of columns of [X]

iii) $H^2 = H$ (show this)

iv) H is symmetric (show this)

(c)

Fact If $\vec{v} \perp \vec{w}$ then

$$\|\vec{v}+\vec{w}\|^2 = \|\vec{v}\|^2 + \|\vec{w}\|^2$$

Proof: $\|\vec{v}+\vec{w}\|^2 = \langle \vec{v}+\vec{w}, \vec{v}+\vec{w} \rangle$

$$= (\vec{v}+\vec{w})' (\vec{v}+\vec{w})$$

$$= (v' + w') \cdot (v+w)$$

$$= v'v + \underbrace{v'w}_{0} + \underbrace{w'v}_{0} + w'w$$

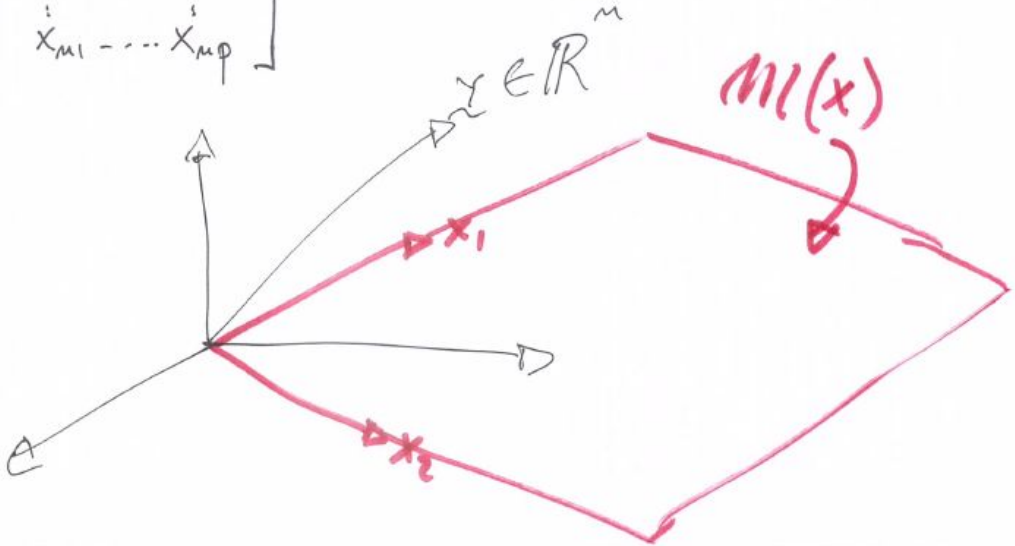
$= \|v\|^2 + \|w\|^2$ pois são $\perp$'s

Assumption: Column $\underset{\sim}{1}$ is in X (d)

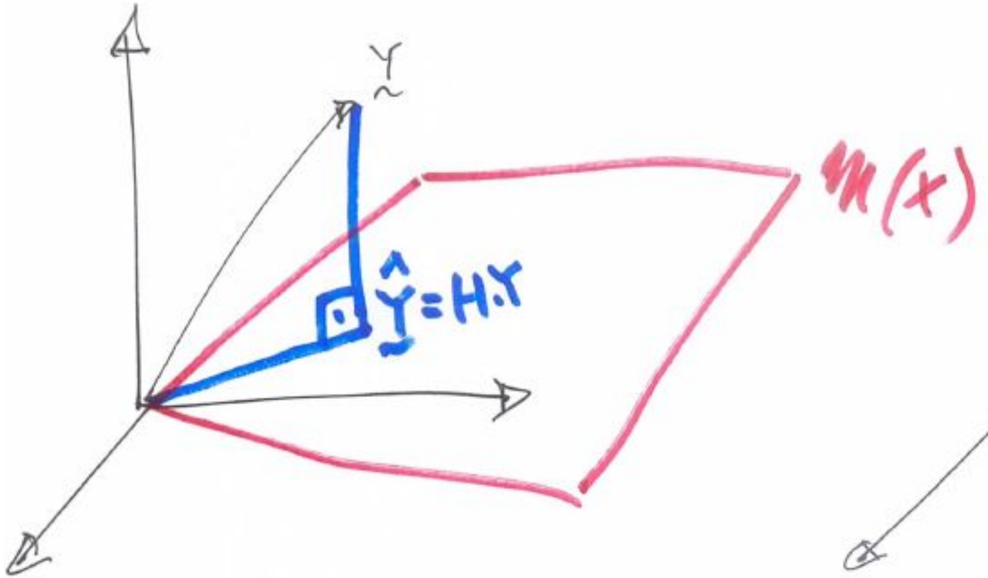
$$\underset{n\times(p+1)}{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

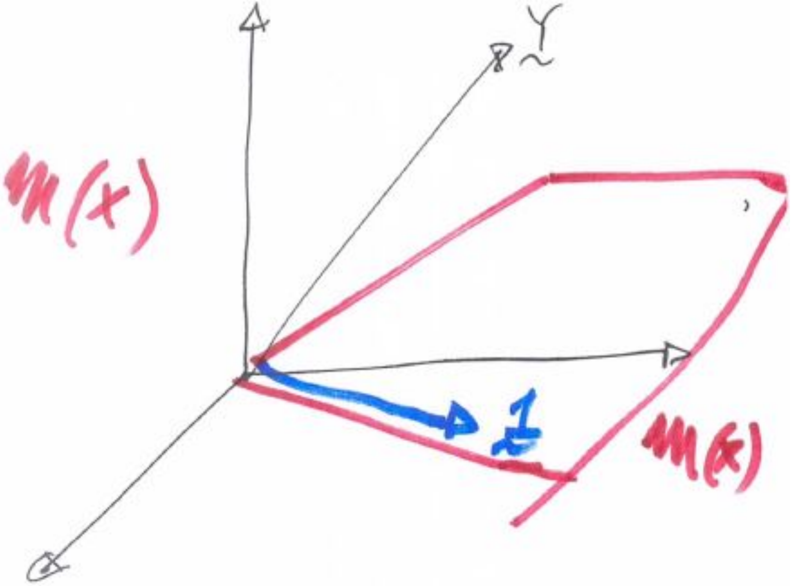
$$\Rightarrow \underset{\sim}{1} \in \mathcal{M}(X)$$

$$\underset{\sim}{1} = 1\cdot\underset{\sim}{1} + 0\cdot\underset{\sim}{x}_1 + \ldots + 0\cdot\underset{\sim}{x}_p$$

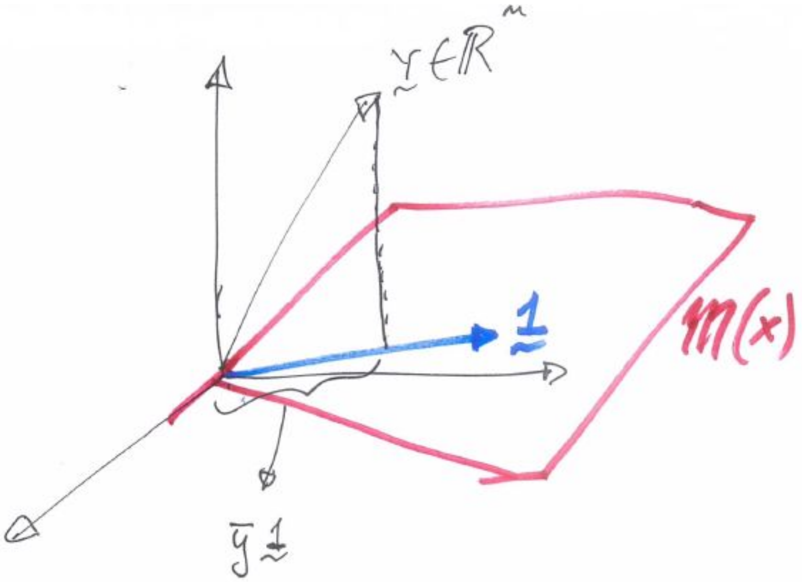


(e)





(f)



$$\bar{y}\,\underset{\sim}{1} = \left(\frac{1}{n}\sum y_i\right)\underset{\sim}{1} = \frac{1}{n}\underbrace{\left(\underset{\sim}{1}'\cdot\underset{\sim}{Y}\right)}_{\substack{\parallel \\ (1,\dots,1)\cdot\begin{pmatrix} y_1 \\ \vdots \\ y_n\end{pmatrix}}}\cdot\underset{\sim}{1} = \frac{1}{n}\,\underset{\sim}{1}\,\underbrace{\left(1'\cdot Y\right)}_{\text{é um escalar}}$$

$$= \frac{1}{n}\left(\underset{\sim}{1}\cdot\underset{\sim}{1}'\right)\cdot Y$$

$$= U\cdot Y$$

↳ matriz associada com $\underset{\sim}{1}$

→ U é matriz de projeç $\perp$ no vetor $\underset{\sim}{1}$ (9)

Prova: basta mostrar que, para todo $Y \in \mathbb{R}^n$, temos $(UY) \perp (Y - UY)$.

$$(UY)' \cdot (Y - UY) = (Y'U') \cdot ((I-U) \cdot Y) = Y' \left(U'(I-U) \right) \cdot Y$$

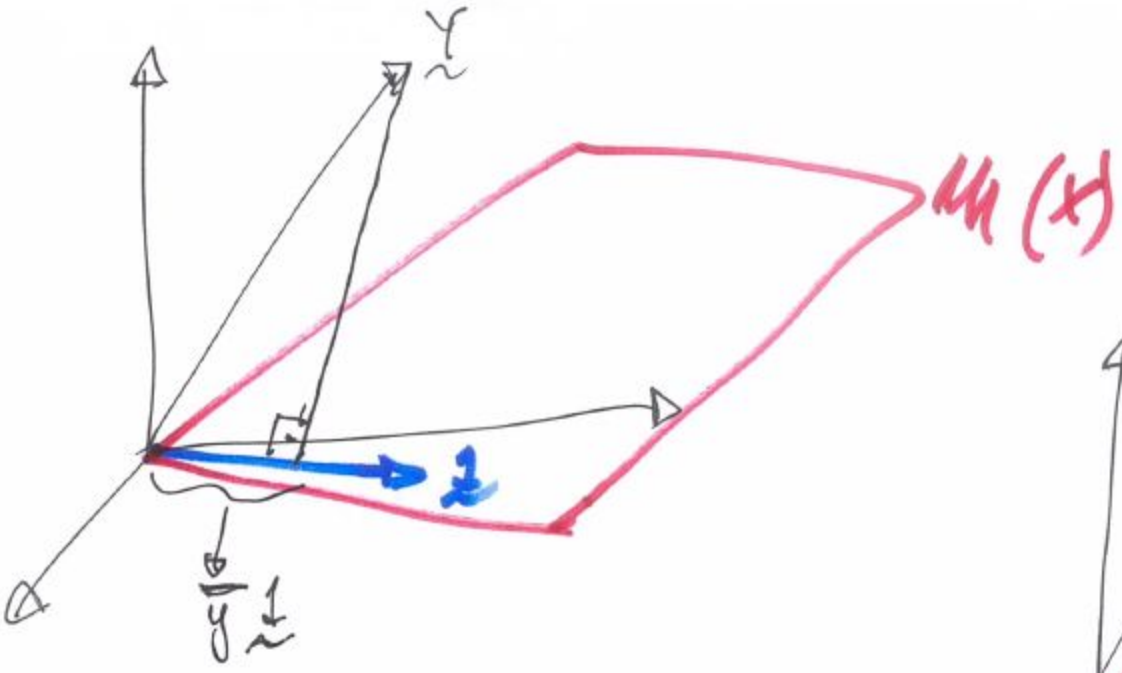
$$= Y' \left(U' - U'U \right) Y.$$

Mas $U' = U$ (show this) $\Longrightarrow$ $U' - U'U = U - U^2$

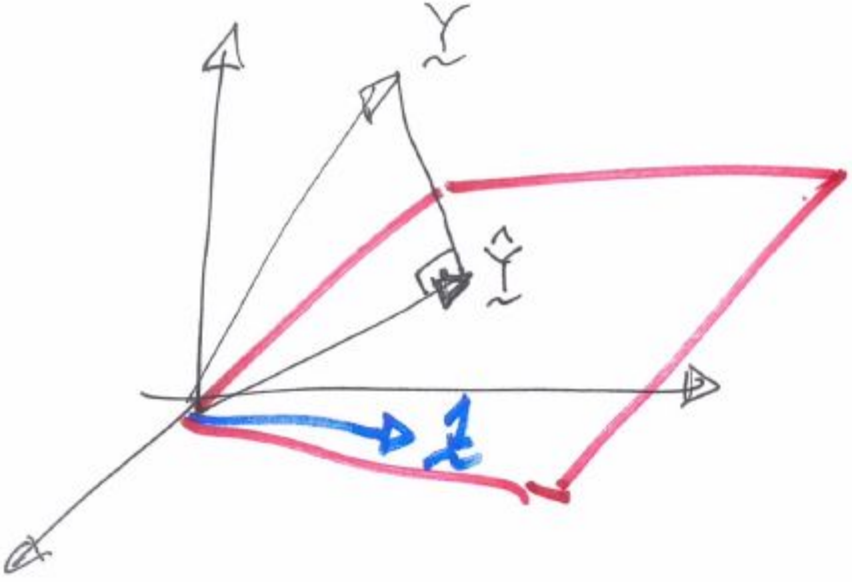
$U^2 = U$ (show this) $= 0$

q.e.d

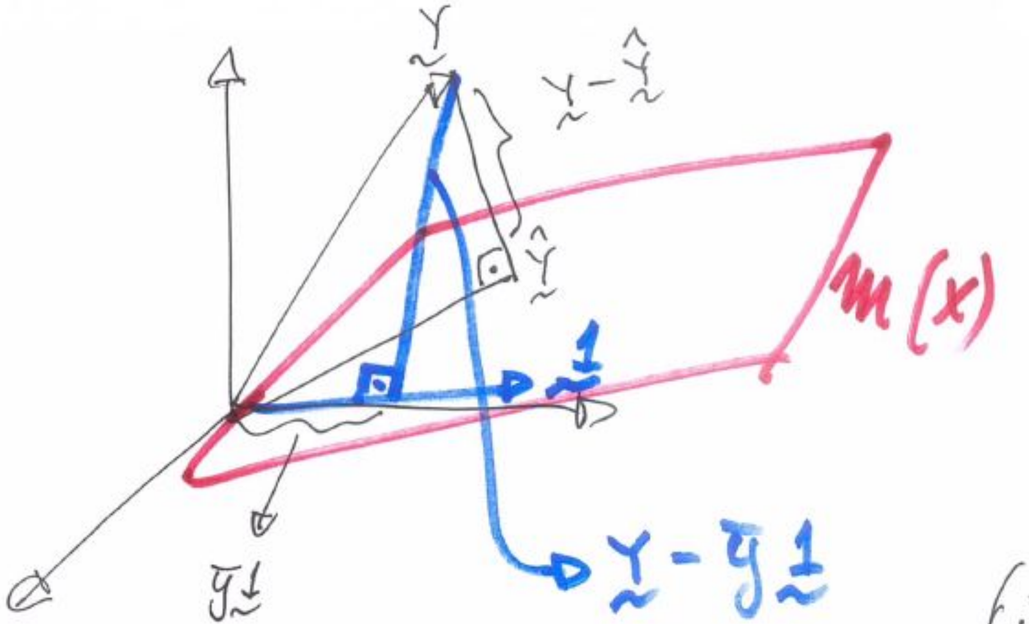
h



$\underset{\sim}{Y}$
$M(X)$
$\underset{\sim}{1}$
$\bar{y}\,\underset{\sim}{1}$



$\underset{\sim}{Y}$
$\underset{\sim}{\hat{Y}}$
$\underset{\sim}{1}$



$$\underset{\sim}{Y} - \bar{y}\,\underset{\sim}{1} = \underbrace{\hat{\underset{\sim}{Y}} - \bar{y}\,\underset{\sim}{1}} + \underbrace{\underset{\sim}{Y} - \hat{\underset{\sim}{Y}}}$$

(identifique $\hat{\underset{\sim}{Y}} - \bar{y}\underset{\sim}{1}$ na figura) d

$$\underset{\sim}{Y} - \bar{y}\,\underset{\sim}{1} = \left(\hat{Y} - \bar{y}\,\underset{\sim}{1}\right) + \left(\underset{\sim}{Y} - \hat{\underset{\sim}{Y}}\right)$$

$\perp$'s !!!

$$\underset{\sim}{Y} - U\underset{\sim}{Y} = \left(H\underset{\sim}{Y} - U\underset{\sim}{Y}\right) + \left(\underset{\sim}{Y} - H\underset{\sim}{Y}\right)$$

$$= \underbrace{(H-U)\cdot\underset{\sim}{Y}} + \underbrace{(I-H)\cdot\underset{\sim}{Y}}$$

$\perp$??

Calcular $\left\langle (H-U)\cdot\underset{\sim}{Y}\,,\ (I-H)\cdot\underset{\sim}{Y}\right\rangle = (*)$

$$(*) = \left((I-H)Y\right)' \cdot (H-U)\,\underset{\sim}{Y} = \underset{\sim}{Y}'\,(I-H)'\cdot(H-U)\,\underset{\sim}{Y} =$$

$$= \underset{\sim}{Y}'\left((I-H)\cdot(H-U)\right)\cdot\underset{\sim}{Y} = Y'\left(\underbrace{H - U - H^2}_{=\text{'s}} + HU\right)\underset{\sim}{Y}$$

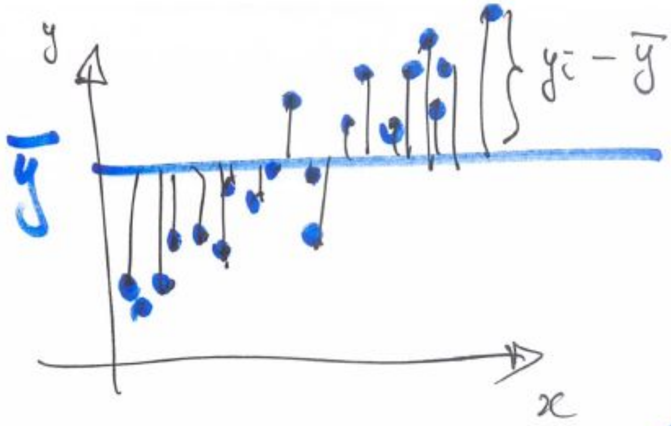
$$= Y'\left(-U + HU\right)\cdot Y.$$ mas $HU = U \Rightarrow (*) = 0$

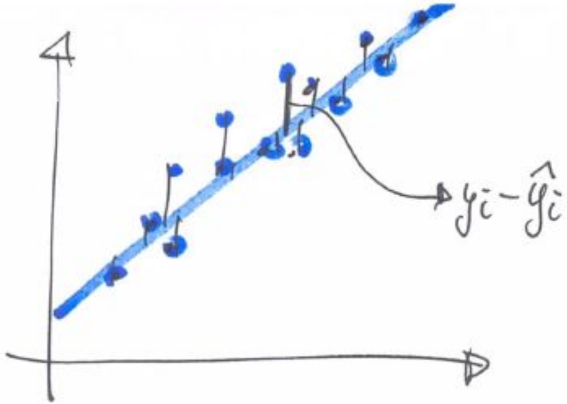
(K)

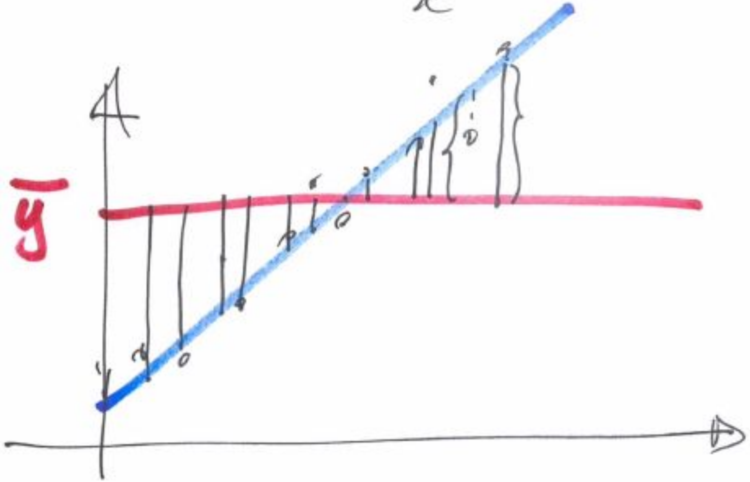
Assim, $(H-U)Y \perp (I-H)Y$ e

$$\underbrace{\| \underset{\sim}{Y} - \bar{y}\, \underset{\sim}{1} \|^2}_{} = \| \hat{\underset{\sim}{Y}} - \bar{y}\, \underset{\sim}{1} \|^2 + \| \underset{\sim}{Y} - \hat{Y} \|^2$$

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{SSTotal} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SSReg} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSRes}$$







$\hat{y}_i - \bar{y}$

(reta $- \bar{y}$)

The sum of squares

- When the residual vector

$$||\mathbf{r}||^2 = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  is small, we have a good fit.
- The idea is to compare this remaining variability with the original variability in $\mathbf{Y}$ BEFORE any regressors were considered.
- The variation of $\mathbf{Y}$ around $\bar{y}$, the mean of $\mathbf{Y}$, is equal to:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = ||\mathbf{Y} - \bar{y}\mathbf{1}||^2$$

Finally, the R^2

- That is, we consider the ratio

$$\frac{||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{||\mathbf{Y} - \bar{y}\mathbf{1}||^2}$$

- If we have a good fit, we should have this ratio close to zero.
- We can prove that this ratio is always smaller than 1.
- Hence, it is more common to use R^2:

$$R^2 = 1 - \frac{||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{||\mathbf{Y} - \bar{y}\mathbf{1}||^2}$$

- A good fit should have $R^2 \approx 1$.

```python
#generate OLS regression results for all features
import statsmodels.api as sm

X_sm = sm.add_constant(X)
model = sm.OLS(y,X_sm)
print(model.fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  csMPa   R-squared:                       0.616
Model:                            OLS   Adj. R-squared:                  0.613
Method:                 Least Squares   F-statistic:                     204.3
Date:                Fri, 15 Oct 2021   Prob (F-statistic):          6.29e-206
Time:                        16:43:15   Log-Likelihood:                -3869.0
No. Observations:                1030   AIC:                             7756.
Df Residuals:                    1021   BIC:                             7800.
Df Model:                           8
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               -23.3312     26.586     -0.878      0.380     -75.500      28.837
cement                0.1198      0.008     14.113      0.000       0.103       0.136
slag                  0.1039      0.010     10.247      0.000       0.084       0.124
flyash                0.0879      0.013      6.988      0.000       0.063       0.113
water                -0.1499      0.040     -3.731      0.000      -0.229      -0.071
superplasticizer      0.2922      0.093      3.128      0.002       0.109       0.476
coarseaggregate       0.0181      0.009      1.926      0.054      -0.000       0.037
fineaggregate         0.0202      0.011      1.887      0.059      -0.001       0.041
age                   0.1142      0.005     21.046      0.000       0.104       0.125
==============================================================================
```

Matriz de projeção ortogonal

Seja X uma matriz de números reais de dimensão $n \times (p+1)$, seja $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ um vetor $(p+1) \times 1$ e $\mathbf{y}$ um vetor $n \times 1$.

Sejam $v_1, \ldots, v_k$ vetores do $\mathbb{R}^n$. Verifique que o conjunto das combinações lineares desses vetores forma um sub-espaço vetorial do $\mathbb{R}^n$.

- Verifique que $X\beta = \beta_0 X_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ onde $X_0, X_1, \ldots, X_p$ são os vetores colunas de X. Assim, o conjunto $\mathfrak{M}(X)$ das combinações lineares das colunas de X é igual a $\mathfrak{M}(X) = \{X\beta \mid \beta \in \mathbb{R}^{p+1}\}$.

Seja X uma matriz de números reais de dimensão $n \times (p+1)$, seja $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ um vetor $(p+1) \times 1$ e $\mathbf{y}$ um vetor $n \times 1$.

- Verifique que $X\beta = \beta_0 X_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ onde $X_0, X_1, \ldots, X_p$ são os vetores colunas de X. Assim, o conjunto $\mathfrak{M}(X)$ das combinações lineares das colunas de X é igual a $\mathfrak{M}(X) = \{X\beta \mid \beta \in \mathbb{R}^{p+1}\}$.

- Seja W um subespaço do espaço vetorial V. Definimos o espaço ortogonal de W como sendo

$$W^{\perp} = \{u \in V \mid \ < u, w >= 0 \ \ \forall \ w \in W\}$$

Mostre que $W^{\perp}$ é um subespaço vetorial de V.
Solução: $\overrightarrow{0} \in W^{\perp}$ pois $< \overrightarrow{0}, w >= 0$ para todo $w \in W$. E tambem $< a_1 u_1 + a_2 u_2, w >= 0$ se $< u_1, w >= 0$ e $< u_2, w >= 0$

A matriz de projeção ortogonal

- Seja $H = X(X'X)^{-1}X'$ de dimensão $n \times n$. Verifique que H é idempotente ($H^2 = H$) e simétrica ($H' = H$).

A matriz de projeção ortogonal

- Seja $H = X(X'X)^{-1}X'$ de dimensão $n \times n$. Verifique que H é idempotente ($H^2 = H$) e simétrica ($H' = H$).
- Seja $y \in \mathbb{R}^n$. A matriz P de dimensão $n \times n$ é dita de projeção ortogonal num certo subespaço vetorial se $y - Py \perp Py$ para todo $y \in \mathbb{R}^n$. Mostre que $H = X(X'X)^{-1}X'$ é uma matriz de projeção ortogonal usando que H é idempotente e simétrica.

A matriz de projeção ortogonal

- Seja $H = X(X'X)^{-1}X'$ de dimensão $n \times n$. Verifique que H é idempotente ($H^2 = H$) e simétrica ($H' = H$).

- Seja $y \in \mathbb{R}^n$. A matriz P de dimensão $n \times n$ é dita de projeção ortogonal num certo subespaço vetorial se $y - Py \perp Py$ para todo $y \in \mathbb{R}^n$. Mostre que $H = X(X'X)^{-1}X'$ é uma matriz de projeção ortogonal usando que H é idempotente e simétrica.

- Como $H = X(X'X)^{-1}X'$ é uma matriz de projeção ortogonal, resta saber em que sub-espaço vetorial W a matriz H projeta os vetores $y \in \mathbb{R}^n$. Mostre que H projeta ortogonalmente em $\mathfrak{M}(X)$ (isto é, mostre que $W = \mathfrak{M}(X)$.)

  Solução: Para todo $y \in \mathbb{R}^n$, temos $Hy = X(X'X)^{-1}X'y = Xb$ onde $b = (X'X)^{-1}\ X'y$. Assim, $Hy \in \mathfrak{M}(X)$ para todo y e portanto $W \subset \mathfrak{M}(X)$. Por outro lado, tome um elemento Xb qualquer de $\mathfrak{M}(X)$. Por definição, $Hy \in W$ para todo y. Em particular, tomando $y = Xb$, temos então $HXb \in W$. Mas $HXb = X(X'X)^{-1}X'Xb = Xb$. Isto é, $Xb \in W$ e portanto $\mathfrak{M}(X) \subset W$. Concluímos então que $W = \mathfrak{M}(X)$.

LS = projeção ortogonal

- Seja $H = X(X'X)^{-1}X'$ a matriz de projeção ortogonal no espaço $\mathfrak{M}(X)$ das combinações lineares das colunas de X. Mostre que ao escolher β tal que $X\beta = Hy$ estamos minimizando a distância $\|y - X\beta\|^2$. DICA: Escreva some e subtraia Hy em $\|y - X\beta\|^2$ e use que $\|v\|^2 =< v, v >$

  Solução: $\|y - X\beta\|^2 =< y - X\beta, y - X\beta >$. Somando e subtraindo Hy obtemos

$$\begin{aligned}
\|y - X\beta\|^2 &= < y - Hy + Hy - X\beta, y - Hy + Hy - X\beta > \\
&= < y - Hy, y - Hy > + < Hy - X\beta, Hy - X\beta > -2 < y - Hy, Hy - X\beta > \\
&= \|y - Hy\|^2 + \|Hy - X\beta\|^2 + 0\,.
\end{aligned}$$

  O último termo acima é zero pois $Hy - X\beta \in \mathfrak{M}(X)$ já que $Hy \in \mathfrak{M}(X)$ e $X\beta \in \mathfrak{M}(X)$ e o conjunto $\mathfrak{M}(X)$ é um sub-espaço vetorial (e portanto contém a diferença dos vetores). Além disso, $y - Hy \in \mathfrak{M}(X)^{\perp}$. Portanto, o produto interno $< y - Hy, Hy - X\beta >$ é nulo.
  Assim, $\|y - X\beta\|^2 = \|y - Hy\|^2 + \|Hy - X\beta\|^2$. O primeiro termo do lado direito não depende de β e o segundo é não-negativo. Ele será minimizado se for igual a zero, o que ocorre se tomamos $X\beta = Hy$.