Projeção ortogonal e mínimos quadrados

Renato Martins Assunção

DCC - UFMG

2017

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 $\mathfrak{M}(X)$.

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

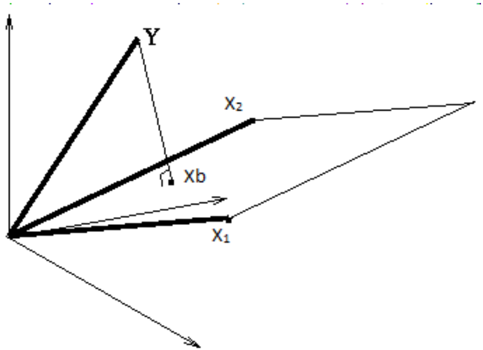

Figura: Representação do vetor $Y \in \mathbb{R}^{1500}$. O plano inclinado representa o sub-espaço vetorial $\mathfrak{M}(X)$ gerado por uma matriz X com apenas duas colunas, os vetores X_1 e X_2, ambos do $\mathbb{R}^{1500}$. O sub-espaço vetorial $\mathfrak{M}(X)$ é de dimensão 2. Identifique visualmente 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:** Leitura opcional, ver no final deste arquivo de slides.

Teorema da Projeção Ortogonal

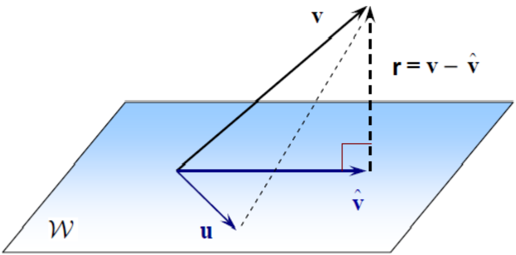

Figura: $\mathbf{v}$ é um vetor do $\mathbb{R}^3$. O plano $\mathcal{W}$ é um sub-espaço vetorial de dimensão 2. dado um vetor $\mathbf{u}$ do sub-espaço, $||\mathbf{v} - \mathbf{u}||$ (linha tracejada fina) é o comprimento do vetor $\mathbf{v} - \mathbf{u}$.
De todos os vetores $\mathbf{u}$ do sub-espaço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

Geometria dos Mínimos quadrados

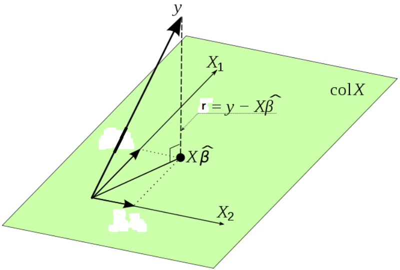

Figura: Projeção ortogonal de $Y \in \mathbb{R}^{1500}$ no sub-espaço vetorial $\mathfrak{M}(X)$ minimiza $||Y - X\beta||^2$. Esta projeção é o vetor $X\hat{\beta}$. Vetor de resíduos é $r = Y - X\hat{\beta}$ e é $\perp$ a $X\hat{\beta}$. Imagem retirada de `https://commons.wikimedia.org/w/index.php?curid=7309159`

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: $\hat{\beta}$ 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 único e é o vetor tal que $X\hat{\beta} \perp (Y - X\hat{\beta})$.
- Em resumo, devemos ter o produto interno zerado: $\langle X\hat{\beta}, (Y - X\hat{\beta}) \rangle = 0$.

Produto interno com vetores-coluna

- Produto interno de dois vetores, **v** e **w**, no mesmo espaço vetorial de dimensão n:

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

- Como **v** e **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 $\mathbf{v}'$.

Encontrando a solução de mínimos quadrados

- Devemos ter

$$\langle\, X\hat{\beta}, (Y - X\hat{\beta})\,\rangle = \hat{\beta}' X^t \cdot (Y - X\hat{\beta}) = 0$$

Isto implca que

$$\begin{aligned} 0 &= \langle\, X\hat{\beta}, (Y - X\hat{\beta})\,\rangle \\ &= \hat{\beta}' X^t Y - \hat{\beta}' X^t X \hat{\beta} \\ &= \hat{\beta}' \left[X^t Y - X^t X \hat{\beta} \right] \end{aligned}$$

- Como isto deve valer para todo Y e X, devemos ter o segundo fator $X^t Y - X^t X \hat{\beta}$ igual a zero.

Encontrando a solução de mínimos quadrados

- Devemos ter
$$\left[X^t Y - X^t X \hat{\beta}\right] = 0$$
ou
$$X^t X \hat{\beta} = X^t Y$$
ou ainda
$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y$$
- A projeção é dada por $\hat{Y} = X\hat{\beta}$.
- Substituindo a expressão acima para $\hat{\beta}$, temos
$$\hat{Y} = X\left(X^t X\right)^{-1} X^t Y.$$

Projeção e predição

- O vetor projetado, $\hat{Y} = X\hat{\beta}$, é valor predito pelo modelo de regressão para cada entrada do vetor Y.
- De fato, em termos matriciais, o vetor $\hat{Y} = X\hat{\beta}$ é igual a

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

Exemplo em Scilab

- Aptos em BH

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,

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

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 $R^2 \approx 1$.

Leitura opcional

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

- Queremos $\hat{Y} = Xb$ tal que $\left(Y - \hat{Y}\right) \perp \hat{Y}$
- Depois, queremos mostrar que este $\hat{Y}$ minimiza $\left|\left|\hat{Y} - Xb\right|\right|^2$

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

- Ou $b^t = 0$ (o que implica que $b = 0$ e que $\hat{Y} = 0$)
- Ou $X^t Y - X^t Xb = 0 \Longrightarrow b = \left(X^t X\right)^{-1} X^t Y$

- Seja $\hat{Y} = X\left(X^t X\right)^{-1} X^t Y = X\hat{b}$
- Vamos calcular agora $||Y - Xb||^2$ em geral:
- Temos $Y - Xb = Y - X\hat{b} + X\hat{b} - Xb$
- Calculando:

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

- Vamos mostrar agora que A =0
$$\left(Y - X\hat{b}\right)^t \left(X\hat{b} - Xb\right) = \left(Y - X\left(X^tX\right)^{-1}X^tY\right)^t \left(X\hat{b} - Xb\right)$$
$$= \left(\left(I - X\left(X^tX\right)^{-1}X^t\right)Y\right)^t \left(X\hat{b} - Xb\right)$$
$$= \left(Y^t\left(I - X\left(X^tX\right)^{-1}X^t\right)^t\right)\left(X\left(\hat{b} - b\right)\right) = (*)$$
- Temos $\left(I - X\left(X^tX\right)^{-1}X^t\right)^t = I^t - \left(X^t\right)^t\left(\left(X^tX\right)^{-1}\right)^t X^t$
$$\begin{aligned} &= I - X\left(\left(X^tX\right)^t\right)^{-1}X^t \\ &= I - X\left(XX^t\right)^{-1}X^t \end{aligned}$$

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

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

- Portanto:
$$||Y - Xb||^2 = \left|\left|Y - X\hat{b}\right|\right|^2 + 0 + \left|\left|X\hat{b} - Xb\right|\right|^2 \geq \left|\left|Y - X\hat{b}\right|\right|^2$$
- Isto é, $X\hat{b}$ minimiza $||Y - Xb||^2$.