# Regressão Linear

Renato Martins Assunção

DCC - UFMG

2015

# Predição de preços imobiliários

- Qual o valor de um imóvel?
- Existem softwares para fazer esta predição de forma automática a partir de várias características do imóvel.
- Menos subjetivo, mais rápido, primeira avaliação.
- Como um software desses pode ser construído?

# Preços de imóveis

- Coletamos preços de 1500 imóveis a venda no mercado de BH.
- Alguns são caros, outros são baratos.
- O que faz com que os preços dos imóveis variem?
- As três coisas mais importantes que afetam o valor de um imóvel...

# Localização

- Localização:
  - Dividir a cidade em pequenas áreas.
- Outra abordagem mais simples:
  - Localização é status socio-econômico;
  - Status é mensurado por renda.
  - Renda é medida pelo IBGE em 2000 pequenas áreas da cidade.
  - Renda do "chefe do domícilio".
- Então: "localização" = renda média da região onde está o imóvel.

# Outras características do imóvel

- Ano da construção
- Área total do imóvel
- Número de quartos
- Número de suítes
- Quantos aptos por andar?
- Possui salão de festas? 0 ou 1
- Possui piscina? 0 ou 1
- ETC...
- Ao todo, 30 características numéricas para cada um dos 1500 imóveis.

# Visão matricial

- Organizar os dados como vetores e matrizes.
- Preços: um vetor $Y$ de dimensão 1500.
- As características: matriz $1500 \times 30$
  - Cada linha = um imóvel
  - 1ª. coluna = renda média da região
  - 2ª. coluna = ano da construção
  - 3ª. coluna = área total
  - Etc.

# Visão matricial

- Preços de 1500 imóveis (vetor de dimensão 1500)

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix}$$

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

- 30 características de 1500 imóveis (Matriz $X$ de dimensao $1500 \times 30$)

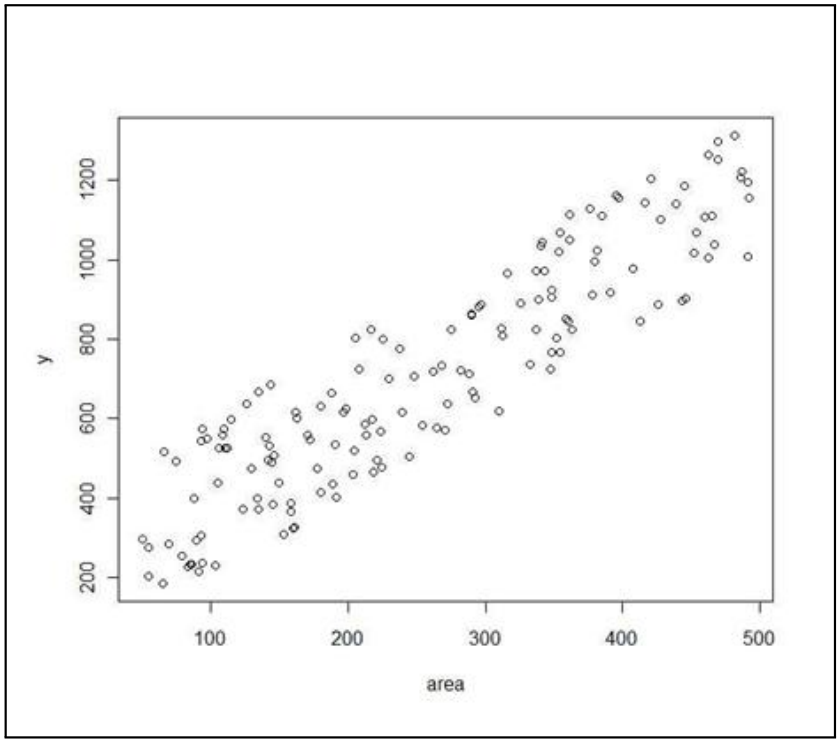
# Preço é uma soma ponderada

- Procuramos um modelo matemático simples que possa explicar, a partir das características, porque alguns imóveis são caros e outros são baratos.
- Área total: quanto maior o imóvel, maior o preço.

# Influência de área

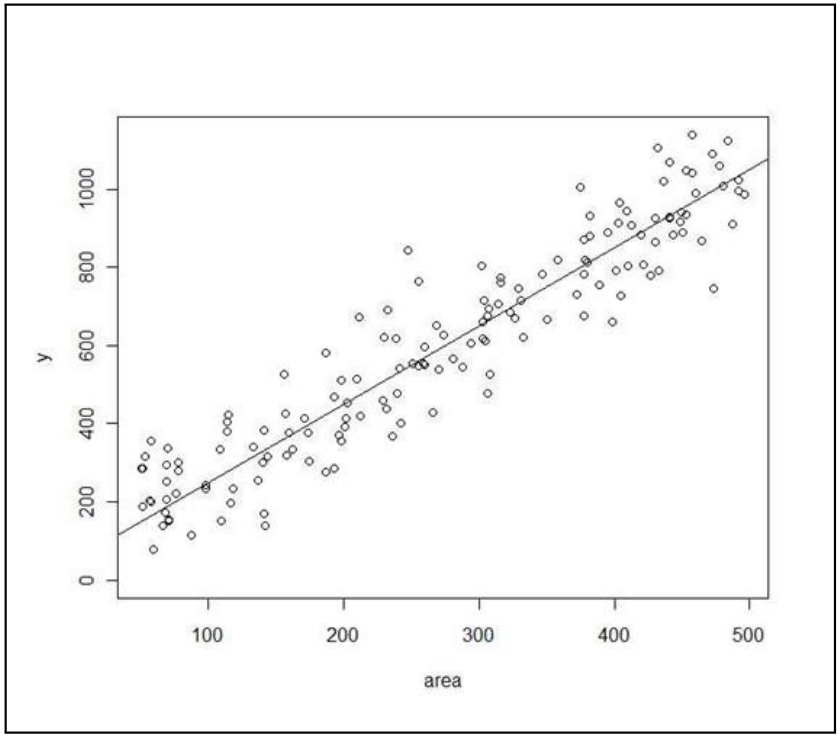
- Vamos fazer uma primeira aproximação, talvez muito grosseira e sujeita a revisões.
- Mas será um ponto de partida.
- Vamos imaginar que, APROXIMADAMENTE, o preço aumenta linearmente com a área do imóvel .
- Isto é, que o preço $Y \approx a + b * \text{área}$.

# Um gráfico com 150 imóveis



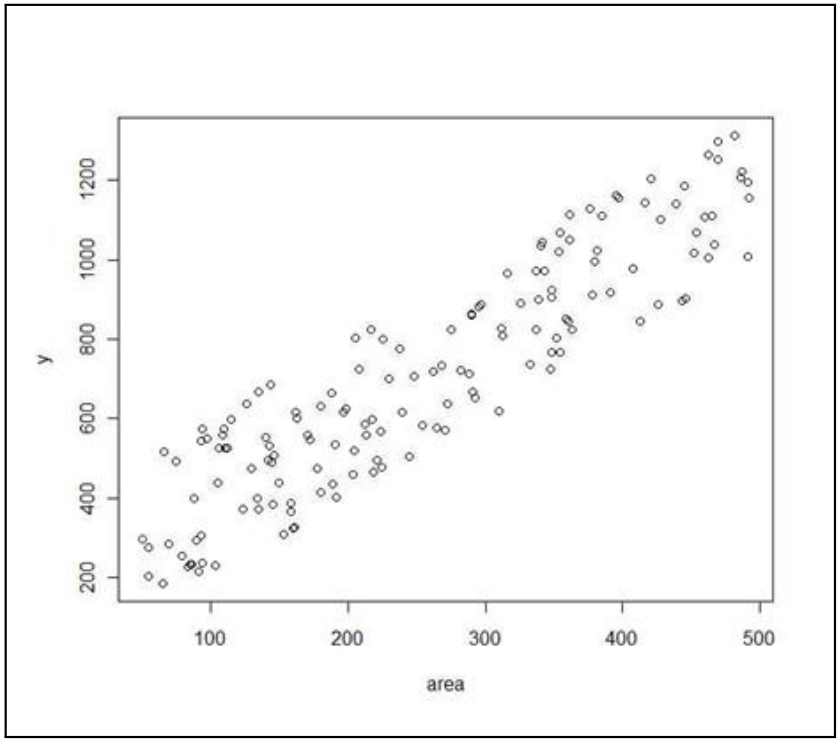

- Cada ponto é um imóvel
  - O eixo vertical tem os preços (em milhares de reais)
  - O eixo horizontal tem as áreas (em metros quadrados)
- Parece que o preço é, grosseiramente, uma função linear da área.
- Isto é, $Y \approx a + b * \text{área}$.

# Um gráfico com 150 imóveis




- Reta no gráfico corresponde a esta equação:
  - Preço
    $Y \approx 50 + 2 * \text{área}.$

# Área não é tudo




- Dois imóveis com praticamente a mesma área possuem preços diferentes.
- O que causa a diferença?
- Idade do imóvel?
- Dois imóveis, com áreas iguais: se um for mais velho, provavelmente será mais barato.

# Ampliando o modelo inicial

- Podemos então imaginar que a idade traz um impacto adicional ao nosso modelo de preço.
- Neste momento, temos $Y \approx a + b * \text{área}$.
- Já vimos até mesmo que $a \approx 50$ e $b \approx 2$
- Podemos agora acrescentar o impacto de idade imaginando que:
  - $Y \approx a + b * \text{área} + c * \text{idade}$.
- Como maior idade reduz o preço, devemos ter $c < 0$.

# Um modelo ainda mais complexo

- Mas o preço não depende apenas de área e idade.
- Dois imóveis com mesma área e mesma idade podem ter preços bem diferentes dependendo de:
  - Sua localização (renda da sua região)
  - Número de suítes
  - Número de vagas na garagem
  - Etc.
- Cada fator pode ser acrescido ao modelo inicial de forma linear.

# Modelo mais complexo

- Vamos considerar um modelo que, a partir das 30 características do imóvel, fornece uma predição do preço da seguinte forma:
  - $Y$ é aproximadamente igual a

$$a + b * \text{área} + c * \text{idade} + d * \text{localização} + \text{ETC} \ldots$$

- O problema é:
  - como encontrar os valores de $a, b, c, etc.$ que tornem a aproximação a melhor possível?

# O problema de forma matemática

- Queremos que cada um desses 1500 valores seja aproximadamente igual a uma combinação linear das 30 características (mais a constante $a$)

$$\begin{aligned} y_1 &\approx a + b * \text{área}_1 + c * \text{idade}_1 + \ldots \\ y_2 &\approx a + b * \text{área}_2 + c * \text{idade}_2 + \ldots \\ &\vdots \\ y_{1500} &\approx a + b * \text{área}_{1500} + c * \text{idade}_{1500} + \ldots \end{aligned}$$

- Podemos escrever isto de forma matricial.

# O problema de forma matemática

- Para facilitar a notação no futuro, vamos escrever os pesos que multiplicam cada característica como $b_0$ (para a constante), $b_1$ (para área), $b_2$ (para idade), ..., $b_{30}$ para a presença ou não de salão de festas

$$
\begin{aligned}
y_1 &\approx b_0 + b_1 * \text{área}_1 + b_2 * \text{idade}_1 + \ldots b_{30} * \text{salão}_1 \\
y_2 &\approx b_0 + b_1 * \text{área}_2 + b_2 * \text{idade}_2 + \ldots b_{30} * \text{salão}_2 \\
&\vdots \\
y_{1500} &\approx b_0 + b_1 * \text{área}_{1500} + b_2 * \text{idade}_{1500} + \ldots b_{30} * \text{salão}_{1500}
\end{aligned}
$$

# O problema de forma matricial

$$
\begin{aligned}
y_1 &\approx b_0 + b_1 * \text{área}_1 + b_2 * \text{idade}_1 + \ldots b_{30} * \text{salão}_1 \\
y_2 &\approx b_0 + b_1 * \text{área}_2 + b_2 * \text{idade}_2 + \ldots b_{30} * \text{salão}_2 \\
&\vdots \\
y_{1500} &\approx \underbrace{b_0 + b_1 * \text{área}_{1500} + b_2 * \text{idade}_{1500} + \ldots b_{30} * \text{salão}_{1500}}
\end{aligned}
$$

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

- Os valores $y_1, y_2, \ldots, y_{1500}$ devem ser colocados em um vetor de dimensão 1500.

# Forma vetorial

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

- $Y$ 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 solução do problema

- Veremos com detalhes mais tarde no curso como resolver este problema.
- Neste momento, basta dizer que nosso problema fica reduzido a um sistema de equações lineares.
- Ou ainda, a um problema de inverter uma certa matriz quadrada.

# A matriz de desenho $X$

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

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

# 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

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

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

Isto é, queremos $Xb \approx Y$. Como resolver isto?

# Solução: um sisteam linear

Queremos encontrar $b$ para resolver o "sistema" linear $Y \approx Xb$

$X$ é uma matriz $1500 \times 31$ e $Y$ é um vetor de 1500 posições.

Como $X$ não é uma matriz quadrada, não é um sistema linear usual: não tem soluç ao, em geral.

# Solução: um sistema linear

Um truque para resolver este "sistema" linear: multiple dos dois lados pela matriz $X^t$ (como se fosse uma constante) e troque $\approx$ por $=$

$$\underbrace{(X^t X)}_{A} b \approx \underbrace{X^t Y}_{c}$$

Assim, terminamos com um sistema linear legítimo do tipo $Ab = c$ onde $A = X^t X$ é matriz quadrada $31 \times 31$ e $c = X^t Y$ é vetor com 31 posições.

# Solução: um sistema linear

- A solução $\mathbf{b} = (b_0, b_1, \ldots, b_{30})^t$ de nosso problema é dada pelo vetor $31 \times 1$ que é a solução desta equação matricial:

$$X^t X \mathbf{b} = X^t Y$$

- Ou ainda, $\mathbf{b} = (X^t X)^{-1} X^t Y$.
- A matriz $X^t X$ é de dimensão $31 \times 1$.
- Inversão via eliminação gaussiana ou, mais profissionalmente, decomposição QR.