# Modelos de Regressão

Renato Martins Assunção

DCC - UFMG

2015

# 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 $1500 \times 31$ 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 $Y \approx Xb$. Como resolver isto?

# Solução: projeção ortogonal

- $X$ é uma matriz $1500 \times 31$, $Y$ 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 $\widehat{b} = \arg\min_b ||Y - Xb||^2$
- Seja $\mathfrak{M}(X)$ o sub-espaço vetorial de $\mathbb{R}^{1500}$ formado pelas combinações lineares das colunas de $X$.
- Se as colunas de $X$ são linearmente independentes, então $\mathfrak{M}(X)$ é um espaço de dimensão igual ao número de colunas de $X$ (que é 31, no nosso exemplo).
- Solução: $Xb$ é a projeção ortogonal de $Y$ em $\mathfrak{M}(X)$.

# Projeção ortogonal

- Seja $W$ um sub-espaço vetorial de $\mathbb{R}^n$.
- A projeção ortogonal de um vetor $Y \in \mathbb{R}^n$ em $W$ é o vetor $w$ tal que $w \perp Y - w$.
- Matriz de projeção ortogonal em $\mathfrak{M}(X)$ é $H = X(X'X)^{-1}X'$
- Para qualquer vetor $Y \in \mathbb{R}^{1500}$, o vetor $HY \in \mathfrak{M}(X)$
- Aleém disso, $HY$ é a projeção ortogonal de $Y$ em $\mathfrak{M}(X)$.
- De fato,

$$HY \,.\, (Y - HY) = Y^t HY - Y^t H^t HY = Y^t HY - Y^t HHY = 0$$

pois $H^t H = HH = H$ (verifique isto você mesmo)

# Solução de mínimos quadrados

- $\widehat{b} = \arg\min_b ||Y - Xb||^2$
- Matriz de projeção ortogonal em $\mathfrak{M}(X)$ é $H = X(X'X)^{-1}X'$
- Para qualquer vetor $Y \in \mathbb{R}^{1500}$, o vetor $HY \in \mathfrak{M}(X)$
- $HY$ é a projeção ortogonal de $Y$ em $\mathfrak{M}(X)$.
- Assim, a solução $Xb$ é $HY = X(X'X)^{-1}X'Y$
- Isto é, $b = (X'X)^{-1}X'Y$

# Uma visão estocástica

- Uma abordagem mais probabilística vai permitir estudar melhor as propriedades do estimador de mínimos quadrados.
- Vamos assumir que o vetor $Y$ é composto de $n$ v.a.'s *independentes.*
- Como estas v.a.'s assumem seus valores altos e baixos?
- Porque alguns aptos tem preços altos e outros possuem preços baixos?
- Vamos explicar como esta variação ocorre quebrando suas causas em dois componentes:
  - Causas determinadas pelos atributos na matriz $X$
  - Outras causas.

# Um modelo para a distribuição de $Y$

- Vamos considerar o $i$-ésimo apto com atributos representado pelo vetor-linha $p$-dim

$$\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{i,p-1})$$

- Procuramos ver o preçco $Y_i$ deste apto aproximadamente como uma função linear dos atributos:

$$Y_i \approx \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p$$

uma combinação dos atributos com pesos FIXOS para todo $i$.

- Vamos agora pensar em $Y_i$ como uma variável aleatória. Vamos escrever

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i$$

- A quantidade aleatória $\varepsilon_i$ é o "erro" aletório de aproximação de $Y - i$ pela função linear dos atributos.

# O erro $\varepsilon_i$

- Considere o "erro" aletório

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p)$$

- Podemos SEMPRE assumir que $\mathbb{E}(\varepsilon_i) = 0$.
- Para ver isto, suponha que $\mathbb{E}(\varepsilon_i) = \alpha \neq 0$.
- Defina um novo erro aleatório $\varepsilon_i^*$ da seguinte forma:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i - \alpha + \alpha \\ &= (\beta_0 - \alpha) + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + (\varepsilon_i - \alpha) \\ &= \beta^* + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i^* \end{aligned}$$

- O 1o. termo do lado direito é uma combinação linear dos atributos
- O novo erro tem

$$\mathbb{E}(\varepsilon_i^*) = \mathbb{E}(\varepsilon_i - \alpha) = \mathbb{E}(\varepsilon_i) - \alpha = \alpha - \alpha = 0$$

# Modelo para $Y_i$

- Assim, temos

$$\begin{aligned} Y_i &= \mathbb{E}(Y_i) + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \end{aligned}$$

- Supomos que os atributos em $\boldsymbol{x}_i$ NÃO SÃO aleatórios.
- Mas, num apto escolhido ao acaso, como supor que o némero de quartos não é uma v.a.?
- Nós assumimos um modelo DISCRIMINATIVO: Condicionamos nos valores dos atributos em $\boldsymbol{x}_i$.
- DADAS AS CARACTERÍSTICAS do apto selecionado em $\boldsymbol{x}_i$, nós supomos que seu preço é

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

# Modelos Discriminativos

- A partir dos $n$ dados, um modelo **discriminativo** aprende a distribuição de probabilidade **condicional** de uma das v.a.'s dadas as demais.
- Vamos isolar uma das variáveis, denotada por $Y$, a variável resposta.
- CONDICIONAMOS em valores específicos e fixos $\boldsymbol{x}' = (x_1, \ldots, y_{p-1})$ para as demais variáveis
- Vamos bolar um modelo para a distribuição de $Y|\boldsymbol{x}$.

# Modelo de regressão linear

- Temos $Y_1, Y_2, \ldots, Y_n$ v.a.'s independentes mas não i.d.
- Como a distribuição muda com $i$?
- Muda em função dos atributos em $\boldsymbol{x}_i'$, a linha $i$ da matriz $X$.
- Até agora temos

$$(Y_i|\boldsymbol{x}_i') = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

- Como nós condicionamos nos valores do vetor de atributos $\boldsymbol{x}_i$, o termo $\boldsymbol{x}_i'\boldsymbol{\beta}$ é um valor não-aleatório.
- Assim, toda a aleatoriedade de $Y_i$ deriva daquela de $\varepsilon_i$:

$$\begin{aligned}
(Y_i|\boldsymbol{x}_i') &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \\
&= \mathbb{E}(Y_i|\boldsymbol{x}_i') + \varepsilon_i \\
&= \text{Termo não-aleatório} + \text{Termo aleatório}
\end{aligned}$$

# Modelo de regressão linear

- Temos $Y_1, Y_2, \ldots, Y_n$ v.a.'s independentes mas não i.d.

$$(Y_i|\boldsymbol{x}_i') = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i = \mathbb{E}(Y_i|\boldsymbol{x}_i') + \varepsilon_i$$

- Já aprendemos que $\varepsilon_i$ é uma v.a. com $\mathbb{E}(\varepsilon_i)$.
- Como $Y_i$ é uma v.a. contínua, então $\varepsilon_i$ também será uma v.a. contínua
- Vamos assumir que $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são i.i.d. $N(0, \sigma^2)$.

# Modelo de regressão linear

- Seja $p(y|\boldsymbol{x})$ a densidade de probabilidade da v.a. $Y$ dados os valores em $\boldsymbol{x}$.
- Os modelos de regressão caem nesta classe de modelos condicionais.
- Queremos um modelo para $Y$ (preço do imóvel) quando são conhecidos os valores de área $(x_1)$ e número de quartos $(x_2)$. Fazemos $\boldsymbol{x} = (x_1, x_2)$.
- Modelo: $(Y|\boldsymbol{x} = (x_1, x_2)) \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$

..

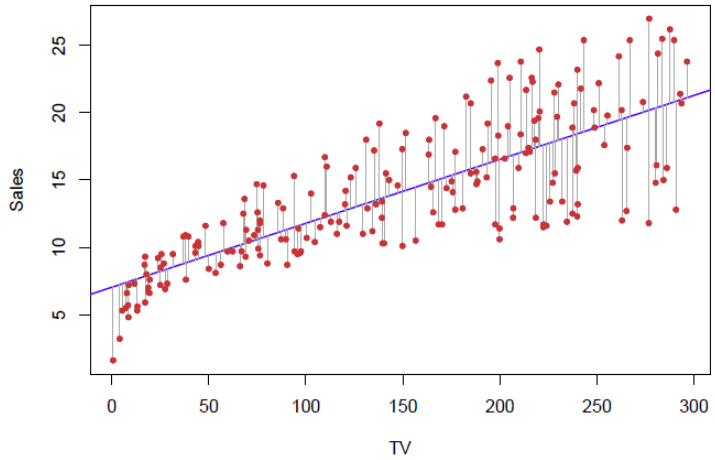


Figura: Modelo de Regressão linear simples: reta "verdadeira" $\beta_0 + \beta_1 x_1$ e dados $(x_{i1}, y_i$ onde $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$.

..

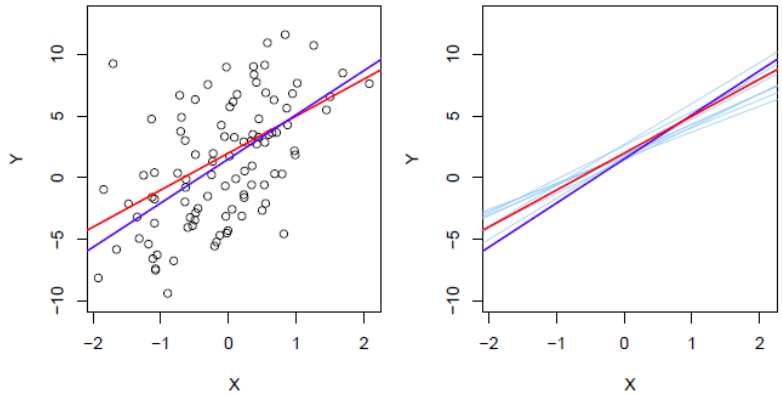


**FIGURE 3.3.** *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, 10 least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Figura: Esquerda: Reta "verdadeira" $\beta_0 + \beta_1 x_1$ e reta estimada $\widehat{\beta}_0 + \widehat{\beta}_1 x_1$ com os dados. Note que $\beta_0 \neq \widehat{\beta}_0$ e que $\beta_1 \neq \widehat{\beta}_1$. Direita: 10 retas estimadas usando 10 diferentes conjuntos de dados, todos gerados independentemente pelo modelo verdadeiro.

# Propriedades do estimador de minimos quadrados

- Estimamos $\beta$ por $\widehat{\beta} = (X'X)^{-1}X'\boldsymbol{Y}$.
- Note que $\widehat{\beta} = AY$ onde $A = (X'X)^{-1}X'$ é uma matriz $p \times n$ com constantes (isto é, uma matriz não-aleatória).
- O vetor $\widehat{\beta}$ é um vetor aleatório porque o vetor $\boldsymbol{Y}$ é um vetor aleatório.
- Como $\boldsymbol{Y} \sim^d N_n(X\beta, \sigma^2 I_n)$ e como $\widehat{\beta} = A\boldsymbol{Y}$ temos

$$\widehat{\beta} = A\boldsymbol{Y} \sim^d N_p\left(A\mathbb{E}(\boldsymbol{Y}), A\Sigma_{\boldsymbol{Y}}A'\right)$$

  (usando as propriedades da normal multivariada)
- Substituindo $A = (X'X)^{-1}X'$ na expressão anterior, nós encontramos

$$\widehat{\beta} = A\boldsymbol{Y} \sim^d N_p\left(\beta, \sigma^2(X'X)^{-1}\right)$$

# Propriedades do estimador de minimos quadrados

- Como

$$\widehat{\boldsymbol{\beta}} \sim^{d} N_p\left(\boldsymbol{\beta}, \sigma^2 (X'X)^{-1}\right) \ ,$$

  temos como consequência que

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$$

- Dizemos que $\widehat{\boldsymbol{\beta}}$ é não-viciado para estimar $\beta$.