Regressão e Mínimos quadrados

Renato Martins Assunção

DCC - UFMG

2016

Lei de Ohm

- Lei de Ohm em circuitos elétricos.
- Resistor cria resistência à passagem de corrente elétrica em circuito.
- Intensidade da corrente depende de características do resistor.
- Corrente depende também da tensão-voltagem aplicada.

Lei de Ohm

- Relação matemática entre tensão (voltagem), corrente e resistência.
- A lei de Ohm:

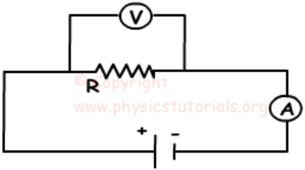
$$V = R\ I$$

- onde:
  - I é a corrente em miliampéres.
  - V é a tensão em volts.
  - R é a resistência em ohms.

  Vamos preferir trabalhar com a lei de Ohm assim:

$$I = \frac{1}{R}\ V$$

Lei de Ohm



If we read the voltage values on voltmeter as V_1, V_2, V_3 and current on ammeter as I_1, I_2 and I_3, there is a relation between them as shown below;

$$\frac{V_1}{I_1} = \frac{V_1}{I_1} = \frac{V_1}{I_1} = \text{constant} = R$$

$$R = \frac{V}{I} \quad \text{or} \quad V = I.R$$

Experimento

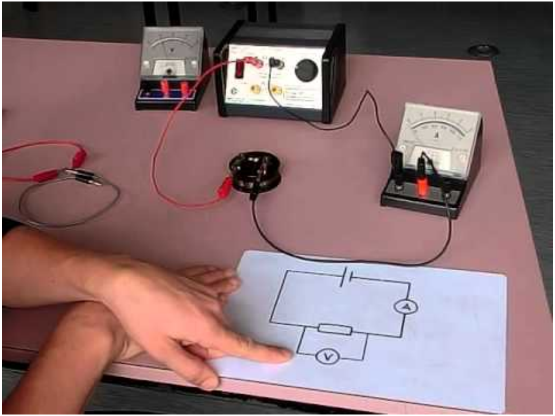


Figura: Montagem do experimento com amperímetro não-digital, com ponteiro.

Dados do Experimento

Voltage	Current
1 V	1 mA
2 V	2 mA
3 V	3 mA
4 V	4 mA
5 V	5 mA
6 V	6 mA
7 V	7 mA
8 V	8 mA
9 V	9 mA
10 V	10 mA

Figura: Dados obtidos com o experimento com uma resistência fixa: altere voltagem e meça corrente resultante.

Gráfico dos dados do experimento

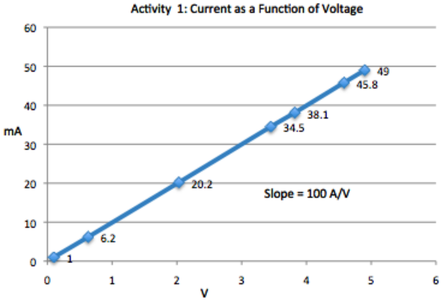


Figura: Um experimento com dados ideais, com lei de Ohm sendo seguida perfeitamente: $I = V/R$. Inclinação da reta $I \times V$ é a resistência R.

Dados de dois experimentos

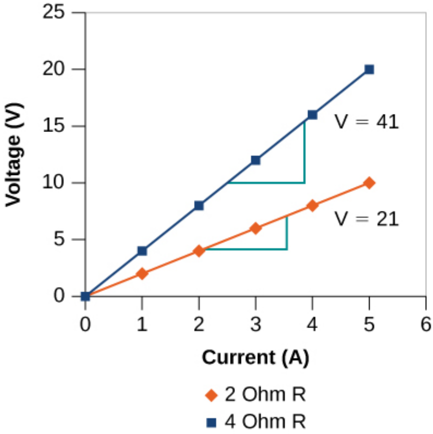


Figura: Dois experimentos com resistores diferentes (eixo trocados em relação à figura anterior). Diferentes resistores implicam diferentes inclinações.

Problema: estimar R

- Num circuito, suponha que não sabemos o valor da resistência R
- Podemos obter uma estimativa para seu valor coletando dados experimentais e usando a lei de Ohm.
- Varie V de 3 a 12 volts: $3, 4, \ldots, 12$
- Vamos denotar $V_1 = 3,\ V_2 = 4, \ldots, V_{10} = 12$

Problema: estimar R

- Para cada valor V_k da voltagem, obtenha o valor correspondente I_k da corrente.
- Faça um gráfico dos pontos (V_1, I_1), $(V_2, I_2), \ldots, (V_{10}, I_{10})$.
- Como $I = \dfrac{1}{R} I$, os pontos devem cair ao longo de uma reta que passa pela origem e com inclinação $1/R$.
- A resistência R desejada é o inverso da inclinação da reta.

Dados reais

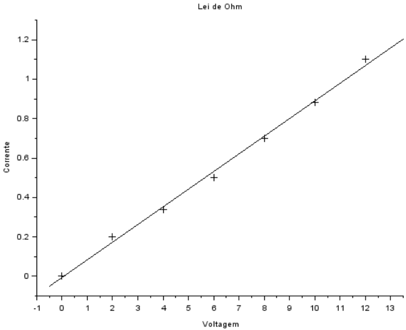


Figura: Dados reais de um experimento não seguem a lei de Ohm $I = V/R$ com perfeição.

Erros em dados reais

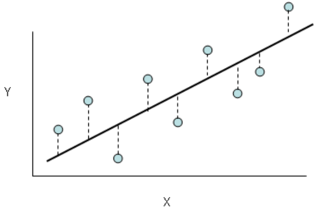


Figura: Dados reais não se alinham perfeitamente ao longo de linha reta por causa de erros de medições.

Dados reais de dois experimentos

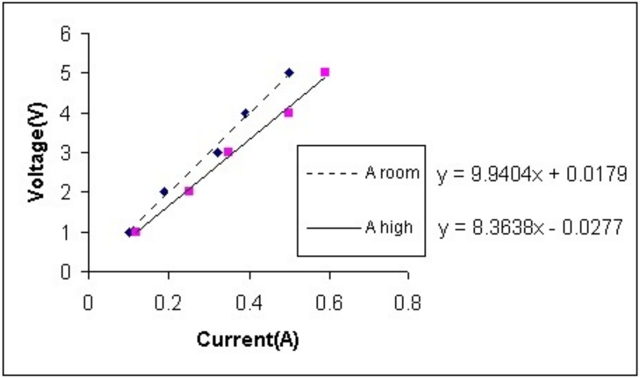


Figura: Dados reais de um experimento não seguem a lei de Ohm $I = V/R$ com perfeição. Dois experimentos com resistores em diferentes temperaturas. Ajuste de mínimos quadrados.

Fonte de imprecisão

- Relação é teoricamente perfeita, mas existem erros de medição.
- Com uma resistência R fixa, faça algumas medições de I e V.
- Elas não seguem a relação $I = \dfrac{1}{R}\ V$ **PERFEITAMENTE**.
- Temos $I \approx \dfrac{1}{R}\ V$
- ou $I = \dfrac{1}{R}\ V + \varepsilon$ onde é um pequeno erro.

Modelo de regressão

- n dados-pontos coletados:

$$(V_1, I_1), (V_2, I_2), \ldots, (V_n, I_n)$$

- Como $I = \dfrac{1}{R} V$, os pontos *deveriam* cair ao longo de uma reta que passa pela origem e com inclinação $\beta_1 = 1/R$.
- Entretanto, temos $I \approx \beta_1 \ V$
- Isto é, cada ponto segue um modelo

$$I_k \approx 0 + \beta_1 V_k$$

ou

$$I_k \approx \beta_0 + \beta_1 V_k + \varepsilon_k$$

- ε_k onde é um pequeno erro em torno de zero.

Dados reais: mostrar em Scilab

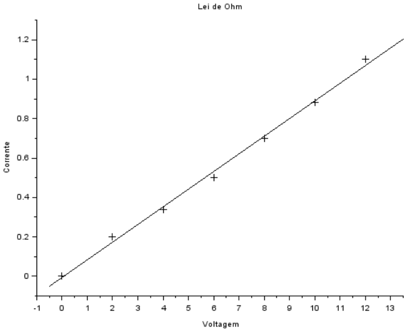


Figura: Dados reais de um experimento: ajuste de regressão de uma reta $\hat{I} = \hat{\beta}_0 + \hat{\beta}_1 V$ como aproximação para o modelo teórico $I = \dfrac{1}{R} V$. Tivemos $\hat{\beta}_0 = -0.0064286$ e $\hat{\beta}_1 = 0.0896429$.

Fonte dos erros

- Com o mesmo aparato experimental (mesma resitência), outros indivíduos medem a corrente com o amperímetro não digital.
- Coletamos mais dados (V, I).
- Vamos plotar os pontos distinguindo os indivíduos.

Dados reais: 4 indivíduos

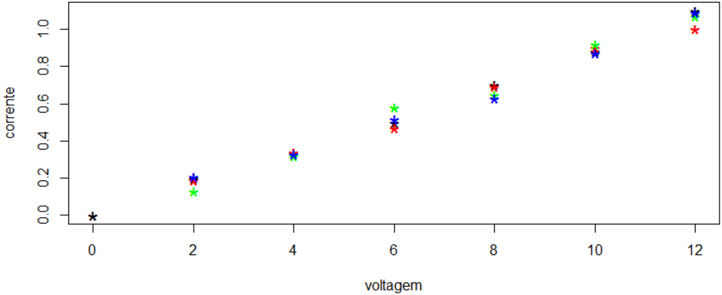


Figura: Dados coletados por 4 indivíduos. Cores distinguem os indivíduos. Mesmas condições e ainda assim, diferentes medições de corrente.

Modelo teórico e aproximação de regressão

- Teoricamente, pela lei de Ohm, temos

$$I = \frac{1}{R}\ V = \beta_0 + \beta_1\ V$$

  onde $\beta_0 = 0$ e $\beta_1 = 1/R$.

- Com dados experimentais, não temos $I = \beta_0 + \beta_1 V$ exatamente.
- Isto é, os pontos (V_k, I_k) não caem exatamente ao longo de nenhuma reta.
- Não caem na reta $I = 0 + \dfrac{1}{R}\ V$.
- Não caem nem mesmo em alguma reta genérica $I = \beta_0 + \beta_1 V$ (possivelmente com $\beta_0 \neq 0$ e $\beta_1 \neq 1/R$). Isto não acontece.

Modelo teórico e aproximação

- Temos $I_k = \beta_0 + \beta_1 \ V_k + \varepsilon_k$.
- Os pontos desviam-se de uma reta teórica.
- O erro ε_k do ponto k pode ser positivo ou negativo.
- Vamos ver um desenho esquemático.

Desenho esquemático

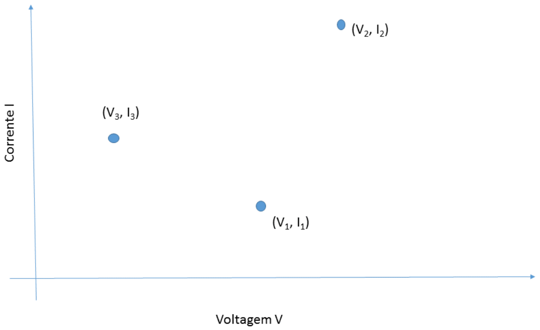


Figura: Dados coletados: 3 pontos (V_1, I_1), (V_2, I_2) e (V_3, I_3). Eles não estão alinhados numa reta.

Desenho esquemático

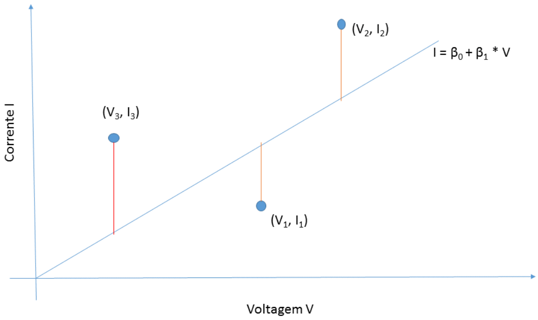


Figura: Dados coletados e linha reta que passa em meio aos dados: uma reta que se ajusta aos dados.

Desenho esquemático

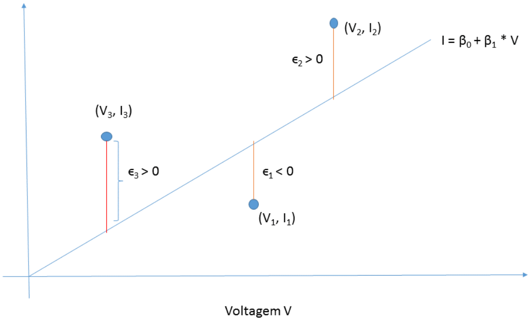


Figura: Os erros ε_1, ε_2 e ε_3. Veja que alguns são negativos e outros positivos.

Problema de regressão

- Temos dados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Eles estão alinhados *aproximadamente* em torno de uma linha reta.
- Isto é,

$$y_k \approx \beta_0 + \beta_1 x_k$$

  para $k = 1, 2, \ldots, n$
- Trocando $\approx$ por $=$:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$$

- Queremos um algoritmo para determinar β_0 e β_1

Casos mais realistas

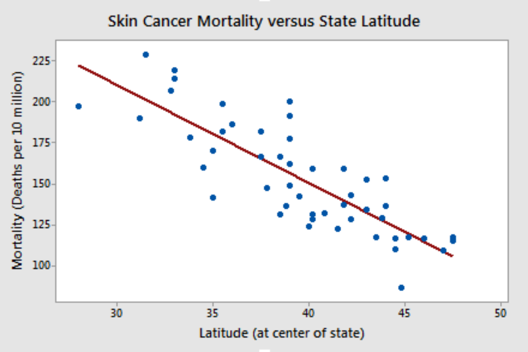


Figura: Cada ponto é uma região. Eixo x: latitude do centro da região. Eixo y: taxa de mortalidade por câncer de pele.

Casos mais realistas

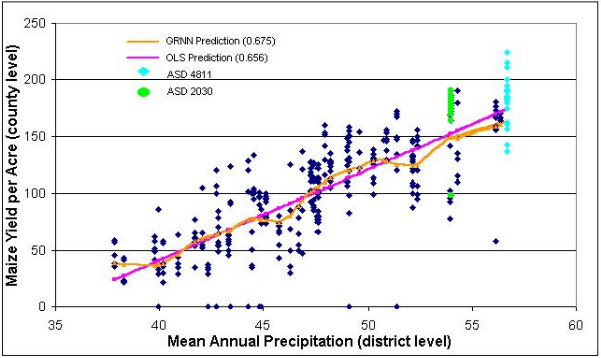


Figura: Cada ponto é uma região. Eixo x: quantidade de chuva. Eixo y:

Casos mais realistas

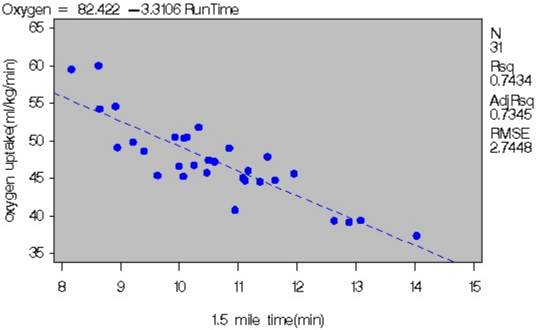


Figura: Cada ponto é um indivíduo após exercícios aeróbicos. Eixo x: tempo correndo numa esteira. Eixo y: Taxa de ingestão de oxigênio.

Casos mais realistas

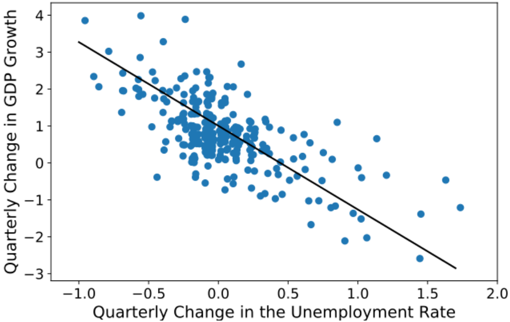


Figura: Lei de Okun em macroeconomia. Cada ponto é um trimestre da economia dos EUA, de 1948 a 2016. Eixo x: mudança percentual de um trimestre para o próximo na taxa de desemprego. Eixo y: mudança percentual no PIB (Gross domestic product, em inglês).
$\Delta GDP \approx 0.789 - 1.654 \times \Delta Desemprego$.

Casos mais realistas

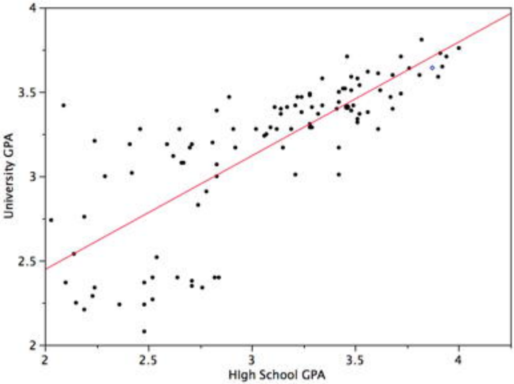


Figura: Cada ponto é um aluno. Notas de alunos no ensino médio e na universidade.

Casos mais realistas

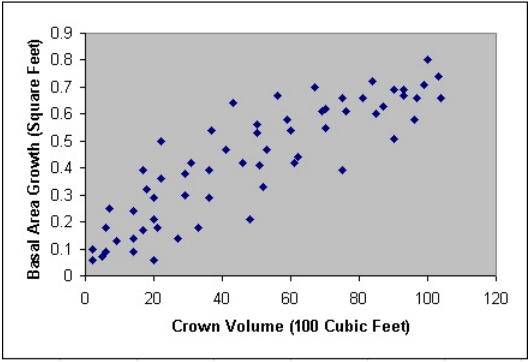


Figura: Cada ponto é uma árvore. Eixo x: área ocupada pelo tronco na base (rente ao chão). Eixo y: área da copa da árvore.

Resumo...

- Temos dados na forma de pontos no plano: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Os dados ficam em torno de uma linha reta:

$$y_k \approx \beta_0 + \beta_1 x_k$$

para $k = 1, 2, \ldots, n$

- Trocando $\approx$ por $=$:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$$

- Queremos um algoritmo para encontrar automaticamente β_0 e β_1.

Em busca de um critério

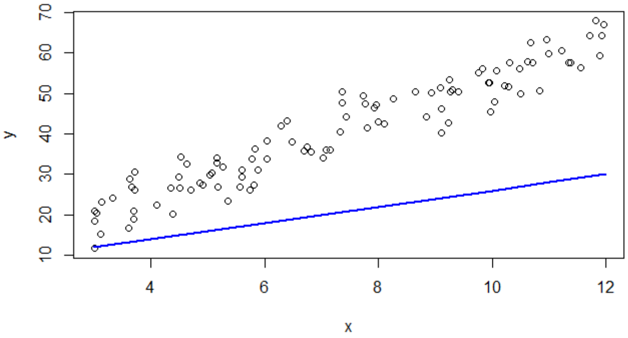


Figura: Um reta tentando se ajustar aos dados: $y = 6 + 2x$. Resultado ruim.

Outra tentativa

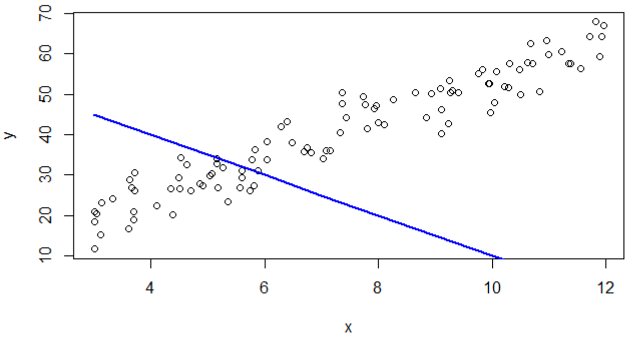


Figura: Outra reta tentando se ajustar aos dados: $y = 60 - 5x$. Resultado pior ainda.

Bom resultado

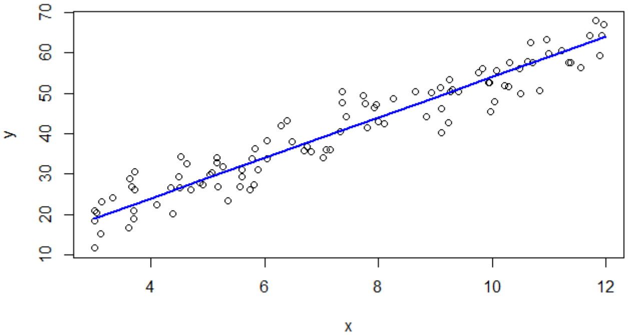


Figura: Um bom ajuste: $y = 4 + 5x$.

Um bom critério

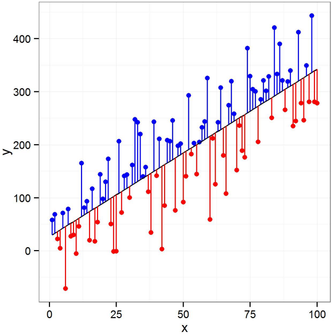


Figura: Reta Candidata: $y = beta_0 + \beta_1 x$. Queremos uma reta tal que os "erros" (segmentos azuis e vermelhos) sejam os menores possíveis.

Uma reta no meio dos pontos

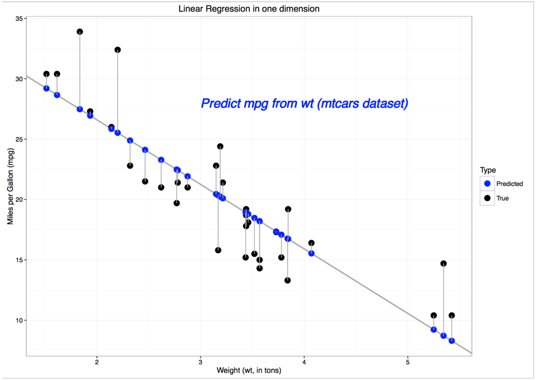


Figura: Reta Candidata: $y = \beta_0 + \beta_1 x$. Para cada ponto (x_i, y_i) obtenha a predição: $\hat{y}_i = \beta_0 + \beta_1 x_i$.

Os resíduos

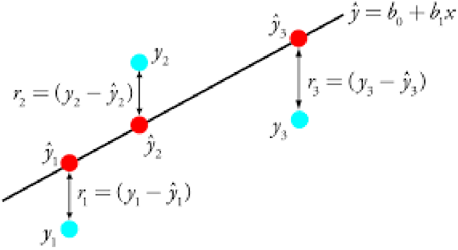


Figura: Obtenha os resíduos: $r_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$.

Uma boa reta minimiza TODOS os resíduos

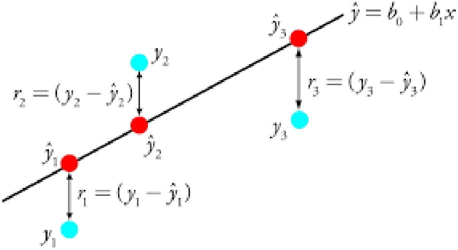


Figura: Não é possível minimizar TODOS os resíduos. Por quê?.

Uma boa reta minimiza a soma dos $|r_i|$

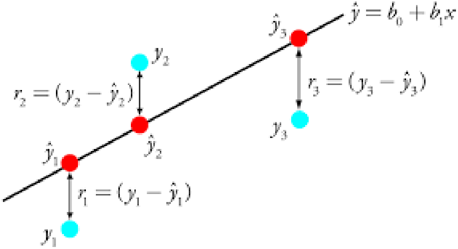


Figura: Ache a reta que minimiza a soma dos resíduos em valor absoluto:

$$\text{minimize} \sum_{i=1}^{n} |r_i| = \sum_i |y_i - (\beta_0 + \beta_1 x_i)|.$$

Não é uma boa ideia...

- Problema: ache a reta (ou β_0 e β_1) que minimiza a soma dos resíduos em valor absoluto:

$$\sum_{i=1}^{n} |r_i| = \sum_i |y_i - (\beta_0 + \beta_1 x_i)|$$

- Veja que temos a soma dos resíduos em valor absoluto é uma função da reta escolhida.
- Para cada reta, temos um conjunto de resíduos r_i e portanto um valor de $\sum_{i=1}^{n} |r_i|$.
- Escrevemos $f(\beta_0 \beta_1) = \sum_i |y_i - (\beta_0 + \beta_1 x_i)|$.

Não é uma boa ideia...

- Problema: ache a reta (ou β_0 e β_1) que minimiza a soma dos resíduos em valor absoluto:

$$f(\beta_0\beta_1) = \sum_{i=1}^{n} |r_i| = \sum_{i} |y_i - (\beta_0 + \beta_1 x_i)|$$

- Queremos achar β_0 e β_1 que minimize $f(\beta_0\beta_1)$.
- Derivamos em β_0 e em β_1 e igualamos o gradiente a zero...
- Mas derivar a função valor absoluto?
- O ponto de mínimo da função $f(r) = |r|$ ocorre em $r = 0$ mas não tem derivada neste ponto.

Outro critério

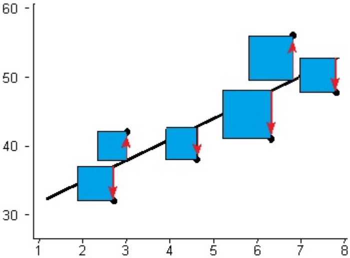


Figura: Uma boa reta minimiza a soma dos r_i^2. Ache a reta que minimiza a soma dos resíduos ao quadrado: minimize $\sum_{i=1}^{n} r_i^2 = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$.

Mínimos quadrados

- Ache a reta (ou β_0 e β_1) que minimiza a soma dos resíduos ao quadrado:

$$f(\beta_0\beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Como encontrar um ponto crítico de $f(\beta_0, \beta_1)$? Derive, iguale a zero e resolva:

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

Mínimos quadrados

- Derivando e igualando a zero:

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

- Ou seja:

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

Mínimos quadrados

- 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.

Equações normais e mínimos quadrados

- Sistema na forma matricial

$$\left[\begin{array}{cc} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{array} \right] \left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right] = \left[\begin{array}{c} \sum_i y_i \\ \sum_i (x_i y_i) \end{array} \right]$$

- Com solução:

$$\left[\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right] = \left[\begin{array}{cc} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{array} \right]^{-1} \left[\begin{array}{c} \sum_i y_i \\ \sum_i (x_i y_i) \end{array} \right]$$

Alguma notação

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

Equações normais com a notação

- 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:

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

Expressão analítica

- 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:

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

Exemplo numérico em Scilab

- Vamos ilustrar o cálculo com um conjunto ridiculamente pequeno de 5 pontos:
- $(23, 77)$, $(22, 53)$, $(88, 160)$, $(65, 170)$, $(31, 74)$.

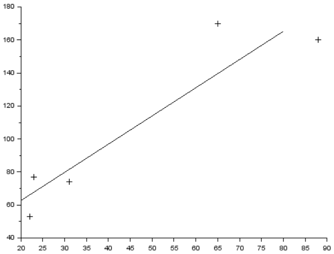


Figura: Exemplo com 5 pontos.

Exemplo numérico em Scilab

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

```
x = [23.   22.   88.   65.   31.]
y = [77.   53.   160.   170.   74.]
num = sum( (x-mean(x)).*(y-mean(y)) )
den = sum( (x-mean(x)).^2 )
b1= num/den // 1.7088688
b0 = mean(y) - b1 * mean(x)   // 28.533808
clf()
plot2d(x, y, style=-1)
plot2d([20,80], b0 + b1*[20, 80])
```

Função reglin em Scilab

```
x=[1.2, 2.5, 4.3, 8.3, 11.6];
y=[6.05, 11.6, 15.8, 21.8, 36.8];
plot(x,y,"o"); // visualizando os dados

// obtendo os coeficientes do ajuste y=a*x+ b usando a funcao
[b1,b0]=reglin(x,y)

clf(); // limpa janela grafica
plot2d([0, 12], b0 + b1*[0, 12]); // grafico da reta
plot(x,y,"o") // acrescentando os pontos
xtitle("Ajuste de regressao linear simples")
```

Regressão múltipla

Regressão múltipla: motivação.

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 trÃªs 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-econÃ´mico;
  - 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 1500×30
  - 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 1500×30)

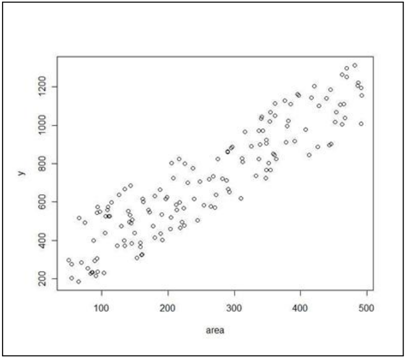
Preço é uma soma ponderada

- 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.

InfluÃªncia de área

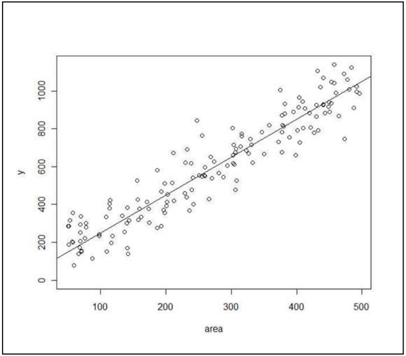
- Vamos fazer uma primeira aproximação, talvez muito grosseira e sujeita a revisões.
- Mas será um ponto de partida.
- Vamos imaginar que, APROXIMADAMENTE, o preço aumenta linearmente com a área do imóvel .
- Isto é, que o preço $Y \approx a + b * \text{área}$.

Um gráfico com 150 imóveis



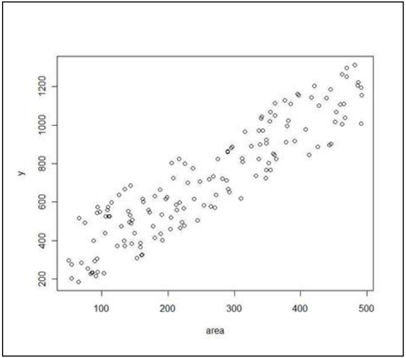
- Cada ponto é um imóvel
  - O eixo vertical tem os preços (em milhares de reais)
  - O eixo horizontal tem as áreas (em metros quadrados)
- Parece que o preço é, grosseiramente, uma função linear da área.
- Isto é, $Y \approx a + b * \text{área}$.

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 matemática

- Empillhe o lado direito de cada uma das 1500 expressões formando um vetor 1500×1:

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

- Ficamos com o vetor 1500×1:

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

O problema de forma matricial

- Escreva o vetor 1500×1 como uma soma de 31 vetores 1500×1:

$$\begin{bmatrix} b_0 + b_1 * \text{área}_1 + b_2 * \text{idade}_1 + \dots b_{30} * \text{salão}_1 \\ b_0 + b_1 * \text{área}_2 + b_2 * \text{idade}_2 + \dots b_{30} * \text{salão}_2 \\ \vdots \\ b_0 + b_1 * \text{área}_{1500} + b_2 * \text{idade}_{1500} + \dots b_{30} * \text{salão}_{1500} \end{bmatrix}$$

- ficamos com

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

- Vamos colocar os valores $y_1, y_2, \dots, y_{1500}$ em um vetor de dimensão 1500.

O problema de forma matricial

- Passamos os coeficientes b_k para fora formando uma combinação linear de 31 vetores 1500×1:

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

- Vamos agora colocar os valores $y_1, y_2, \ldots, y_{1500}$ 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 APROXIMADAMENTE como uma combinação linear de 31 vetores de dimensão 1500.
- Problema: encontrar os 31 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 1500×31 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 sistema linear

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

X é uma matriz 1500×31 e Y é um vetor de 1500 posições.

Como X não é uma matriz quadrada, não é um sistema linear usual: não tem solução, 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 $=$

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

Assim, terminamos com um sistema linear legítimo do tipo $Ab = c$ onde $A = X^tX$ é matriz quadrada 31×31 e $c = X^tY$ 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 31×1 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 31×1.
- Inversão via eliminação gaussiana ou, mais profissionalmente, usando a decomposição QR.

Exemplo em Scilab

```
// Existem 9 colunas de variaveis com os seguintes nomes e de
// lcavol = logaritmo do volume do tumor
// lweight = log(peso da prostata)
// age = log(idade)
// lbph = log(hiperplasia prostatica benigna)
// svi = variavel binaria e indicadora de invasao da veiscula
// lcp = log(penetracao capsular)
// gleason = escore de Gleason, uma nota global associada com
// pgg45 = porcentagem do tumor que pode ser classficado com
// lpsa = log( PSA ) onde PSA = AntÃgeno prostÃ¡tico especÃ-
fico
```

Exemplo em Scilab

```
// O interesse neste estudo e' criar um modelo que sirva para prev
// (a 9a coluna no arquivo) em funcao das outras 8 variaveis.
// Para isto voce vai ajustar um modelo de regressao linear onde
// lpsa = b0 + b1*lcavol + b2*lweight + ... + b8*pgg45
// Leia o arquivo no scilab
M = fscanfMat("prostata.tab");
// crie matriz de desenho X com dimensao 97x8 com as 8 primeiras c
X = M(:,1:($-1)); // $ significa o ultimo indice
y = M(:,$);
```

Exemplo em Scilab

```
[b1,b0]=reglin(X',y') // reglin pede as matrizes na forma transpos
b0  =  0.1813097
b1  = column 1 to 4
      0.5643524   0.622048  -0.0212489   0.0966926
      column 5 to 8
      0.7616526  -0.10605   0.0492518   0.0044577
```