

MINERAÇÃO DE DADOS

Thiago Marzagão¹

¹marzagao.1@osu.edu

REGRESSÃO LINEAR SIMPLES

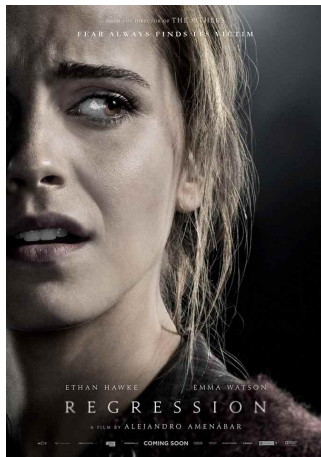
as três grandes áreas da mineração de dados

- regressão
- classificação
- clusterização

as três grandes áreas da mineração de dados

- regressão
- classificação
- clusterização

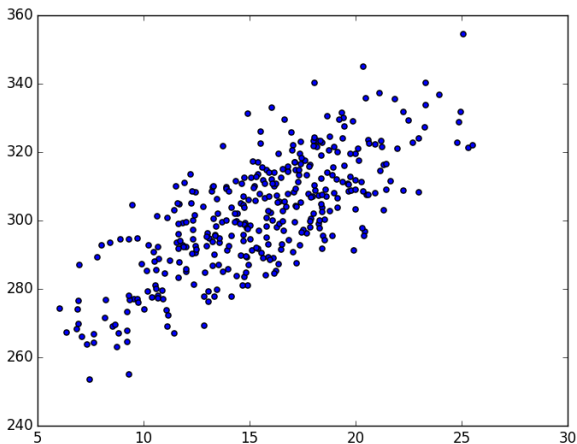
o que regressão *não* é



temperatura vs venda de sorvetes

dia	temperatura	venda de sorvetes
1	17,8°C	R\$ 296,7
2	11,9°C	R\$ 298,3
3	26,8°C	R\$ 323,0
4	12,8°C	R\$ 293,0
5	10,0°C	R\$ 285,7
6	12,7°C	R\$ 287,9
7	18,3°C	R\$ 308,4
8	13,0°C	R\$ 288,6
9	16,9°C	R\$ 321,7
10	14,7°C	R\$ 275,0
11	14,9°C	R\$ 313,7
12	13,8°C	R\$ 276,7
13	10,9°C	R\$ 283,3
...

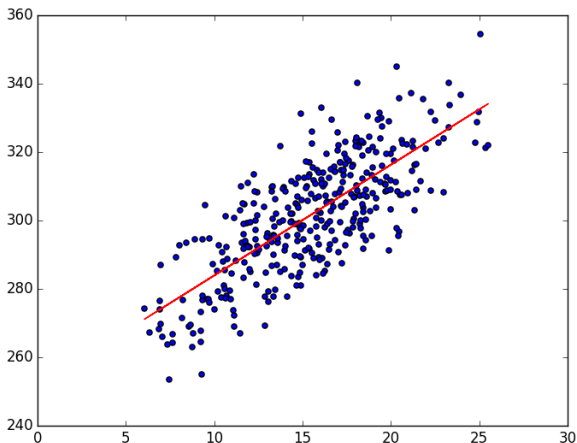
temperatura vs venda de sorvetes



temperatura vs venda de sorvetes

- P/ cada grau centígrado a mais, quantos R\$ a mais de venda?
- Se fizer 25°C amanhã, quantos R\$ de sorvete eu devo esperar vender?

temperatura vs venda de sorvetes



regressão linear: idéia básica

- Se existe uma relação *linear* entre X e Y , então essa relação pode ser modelada como uma reta.
- Equação p/ gerar uma reta: $y = a + bx$
- a e b são constantes
- x e y são variáveis
- Mas como escolher a *melhor* reta?
- Erros absolutos vs erros quadrados (desenhar no quadro).
- P/ o exemplo dos sorvetes: $y = 249,32 + 3,38x$
- x = temperatura (em Celsius)
- y = vendas de sorvete (em R\$)

Mínimos Quadrados Ordinários

- Como encontrar a reta que minimiza os erros quadrados?

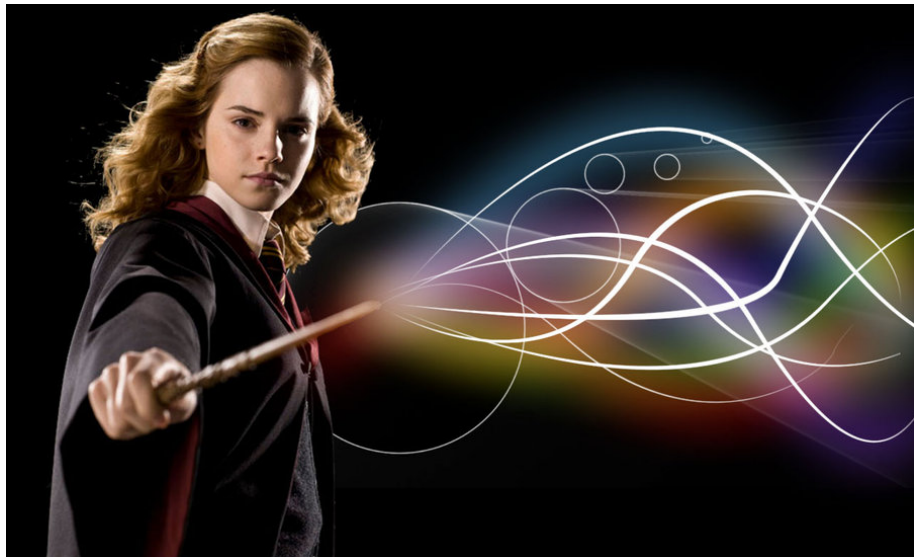
Mínimos Quadrados Ordinários

- Como encontrar a reta que minimiza os erros quadrados?
- Nosso modelo até agora: $y_i = a + bx_i$
- $i = 1, 2, \dots, N$ (N = número de amostras)
- Mas y_i estimado $\neq y_i$ observado.
- Chamemos de \hat{y}_i o valor estimado e de y_i o valor observado.
- Diferença entre valor observado e valor estimado (erro): $e_i = y_i - \hat{y}_i$
- ~~$y_i = a + bx_i$~~
- $\hat{y}_i = a + bx_i$
- $y_i = a + bx_i + e_i$
- $e_i = y_i - a - bx_i$
- $\min_{a,b} \sum_{i=1}^N e_i^2 = \min_{a,b} \sum_{i=1}^N (y_i - a - bx_i)^2$
- Em português: queremos encontrar o a e b que minimizam a soma dos erros quadrados.

Mínimos Quadrados Ordinários

- $\min_{a,b} \sum_{i=1}^N e_i^2 = \min_{a,b} \sum_{i=1}^N (y_i - a - bx_i)^2$
- P/ encontrar a solução é preciso calcular as derivadas parciais:
- $$\frac{\partial \sum_{i=1}^N (y_i - a - bx_i)^2}{\partial a} = \sum_{i=1}^N -2(y_i - a - bx_i) = 0$$
- $$\frac{\partial \sum_{i=1}^N (y_i - a - bx_i)^2}{\partial b} = \sum_{i=1}^N -2x_i(y_i - a - bx_i) = 0$$

a magia do cálculo multivariado acontece...



Mínimos Quadrados Ordinários

- $b = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$
- $a = \bar{y} - b\bar{x}$
- \bar{x} é a média de x
- \bar{y} é a média de y

Mínimos Quadrados Ordinários

- Como interpretar a e b ?
- b é a variação em y quando x varia uma unidade
- Exemplo dos sorvetes: $\bar{y}_i = 249,32 + 3,38x_i$
- P/ cada 1°C a mais de temperatura, a venda de sorvetes aumenta em R\$ 3,38.
- P/ cada 1°C a menos de temperatura, a venda de sorvetes diminui em R\$ 3,38.
- O sinal importa! a e b podem ser positivos ou negativos.
- a é o valor de y quando $x = 0$
- Quando a temperatura é de 0°C , a venda de sorvetes é de R\$ 249,32.

Mínimos Quadrados Ordinários

- x causa y ?
- Não!

Mínimos Quadrados Ordinários

THE DEADLY FACTS ABOUT WATER!

FACT!

WATER CAN BE CHEMICALLY
SYNTHESIZED BY BURNING
ROCKET FUEL!!!

FACT!

OVER CONSUMPTION CAN CAUSE
EXCESSIVE SWEATING, URINATION,
AND EVEN DEATH!!!

FACT!

100%
OF ALL SERIAL KILLERS,
RAPIST AND DRUG DEALERS HAVE
ADMITTED TO DRINKING WATER!!!



FACT!

WATER ONE OF THE PRIMARY INGREDIENTS
IN HERBICIDES AND PESTICIDES!!!

FACT!

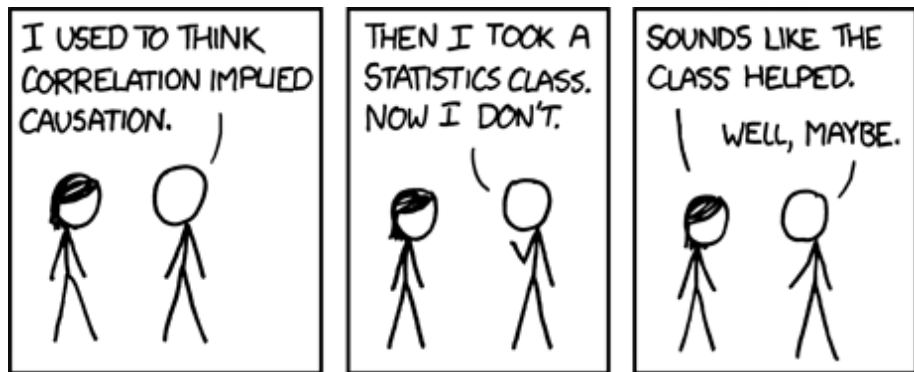
WATER IS THE LEADING
CAUSE OF DROWNING!!!

FACT!

100 PERCENT OF ALL PEOPLE
EXPOSED TO WATER WILL DIE!

Mínimos Quadrados Ordinários

- b nos diz apenas a variação esperada em y dada uma variação em x
- b não nos permite dizer que x causa y



Mínimos Quadrados Ordinários

- E quando x não representa uma quantidade?
- Digamos, $x = \text{choveu/não choveu}$.
- $x = \text{homem/mulher}$
- $x = \text{município de residência}$
- etc
- Nesses casos é preciso codificar x
- Ex.: $x = 0$ se não choveu, $x = 1$ se choveu
- Venda de sorvetes vs chuva: $\hat{y}_i = a + bx_i$
- Não choveu: $\hat{y}_i = a + b(0) = a + 0 = a$
- Choveu: $\hat{y}_i = a + b(1) = a + b$
- Nesses casos x é chamado de variável dummy.
- p / n categorias, crie $n - 1$ dummies

Mínimos Quadrados Ordinários

- E quando y não representa uma quantidade?
- Digamos, $y = \text{choveu/não choveu}$.
- Aí é um problema de classificação, não de regressão.

- Múltiplos x s: $\hat{y}_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + \dots$

o que *não* vamos ver

- Séries temporais.
- Dados em painel.
- Inferência (b é estatisticamente diferente de zero).
- Causalidade (x causa y ? y causa x ? z causa x e y ?).
- Em mineração de dados o foco geralmente é na predição: dados x_1 , x_2 , etc, queremos prever (ou classificar) y .