Data Mining
Análise por Regressão II

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2019

www.inf.ufsm.br/~joaquim

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

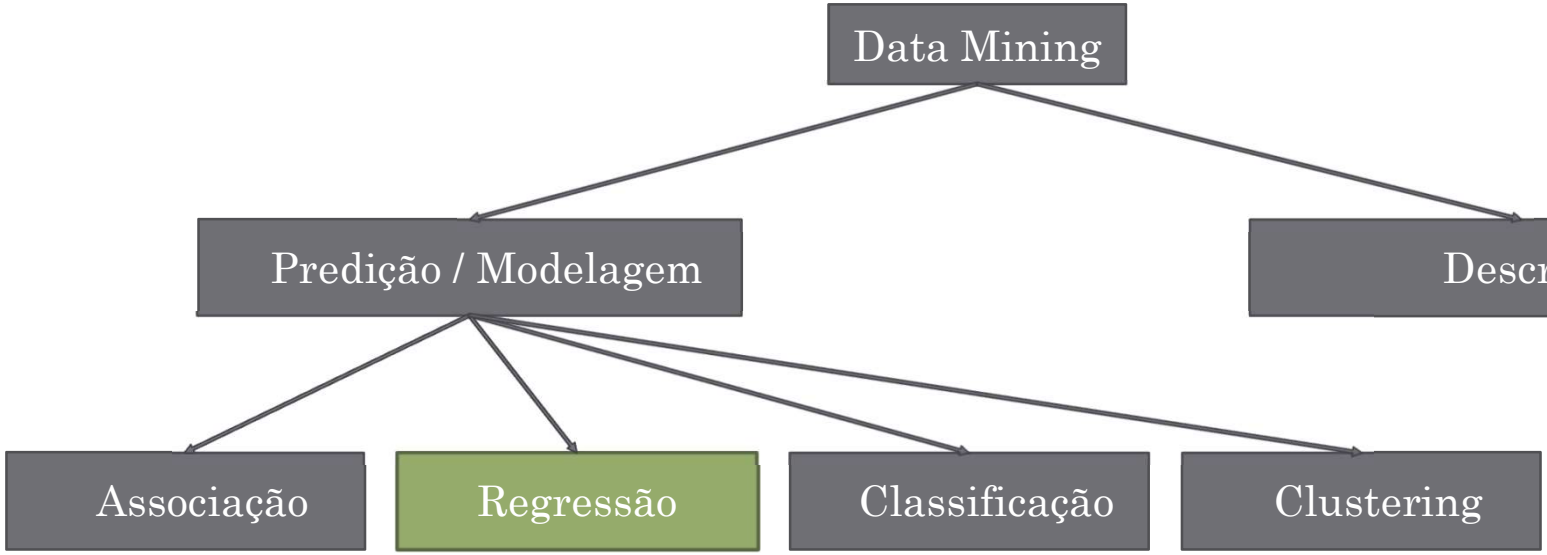
Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

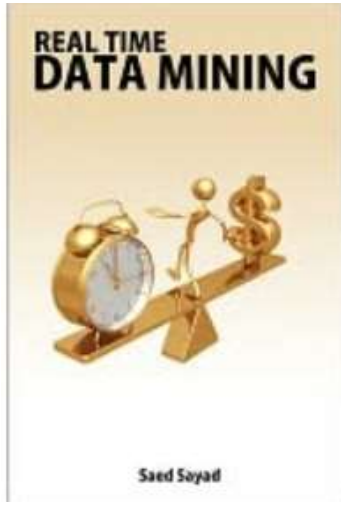
*Caso você queira usar algo desse material em alguma publicação, envie-me um e-mail com antecedência.

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Mapa para Mineração de Dados*





***http://www.saedsayad.com/data_mining_map.htm**

Vamos começar

1. Abra o dataset `galton` da biblioteca "`UsingR`".
2. Crie um novo set `galton.br` com as variáveis traduzidas (use `names()`).
3. Crie um histograma para os filhos passando "breaks=70" como parâmetro.

Centro da massa

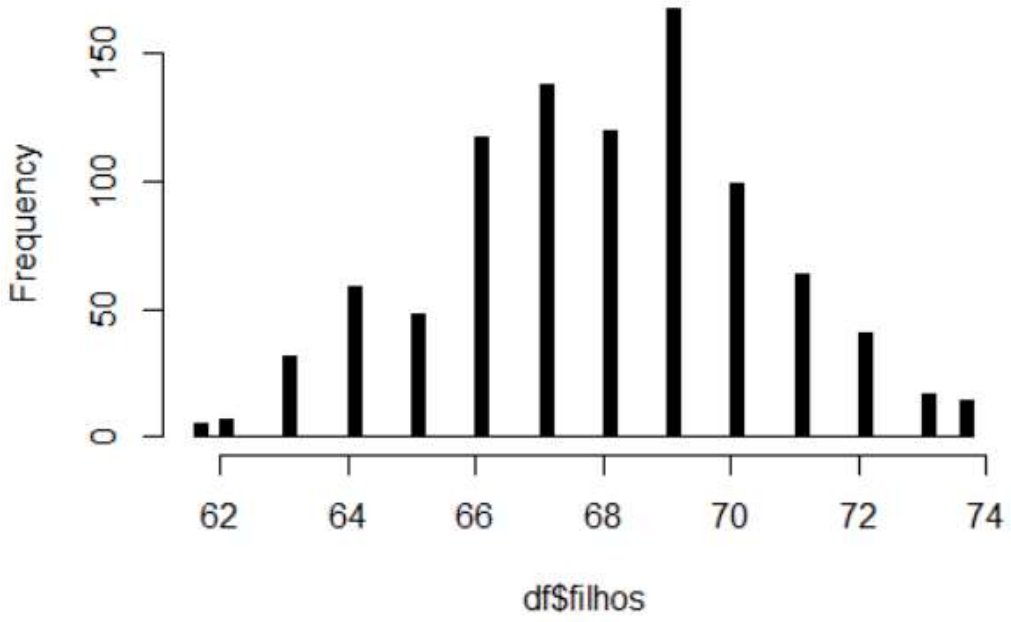
- O centro da massa de um histograma é descrito pela variável μ que minimiza a seguinte equação:

1/n = $\sum_{i=1}^{n}(Y_i - \mu)^2$

- Sendo `Y` um conjunto de `n` filhos.
- Este será o centro físico da massa.

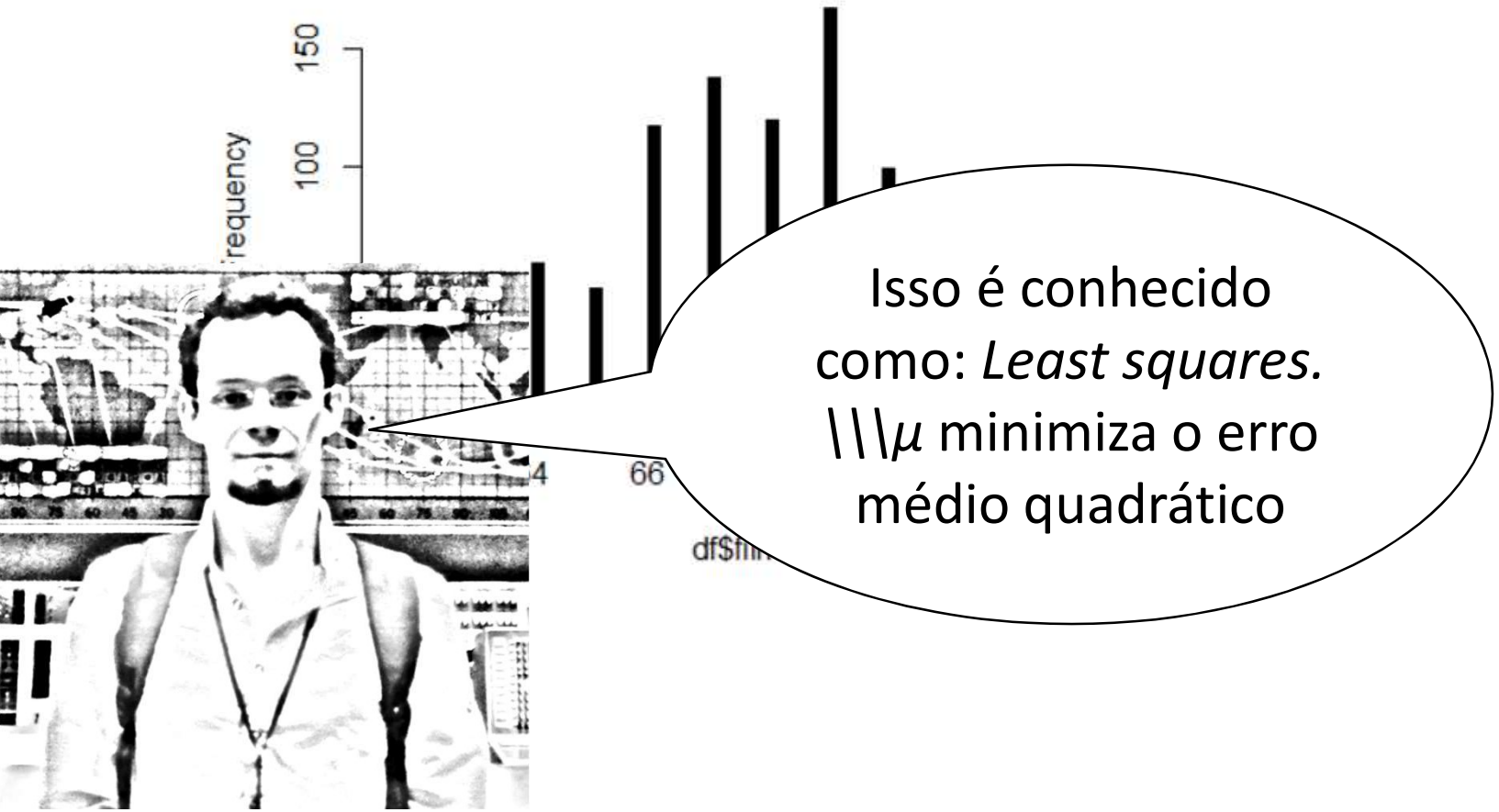
…Como calcular \mu de forma fácil?

Centro da massa



…Imagine uma objeto que você tem que equilibrar com uma mão. Que parte do eixo X voc6e deve segurar para que o objeto não caia?

Centro da massa



Hands On!

- Crie um histograma (`hist()`)para mostrar a altura média dos filhos (`galton` dataset). Use a biblioteca `manipulate` com um range de 160 a 175 e um passo de 0.7 para ver a evolução do erro quadrático. Descubra o melhor μ de acordo com o erro quadrático médio. Mostre o erro quadrático médio.

Technical help

Exemplo de função em R

```
func <- function(parametro){ #faca algo }
```

Use `manipulate(minhaFuncao(meuParametro), parametro=slider(ini,fim, step=X))`

Use `lines` e `text` para, respectivamente, colocar uma linha em seu gráfico e imprimir um texto

Regressão para a média

- Considere que a origem foi fixada (0,0). Qual a melhor linha regressiva para representar os dados?
- A linha que corta o plano cartesiano possui um ângulo ß que ao ser ajustado minimiza a distância dos demais pontos.
- ?! Isso facilita nossa vida, pois podemos pensar em um único parâmetro!

Regressão para a média

- Porém, não faz muito sentido começar do (0,0) porque a linha regressiva nunca vai chegar no seu melhor estado.
- A melhor estratégia, nesses casos, é começar pela média. Logo $x \leftarrow serie - media(serie)$

Hands On!

- Vamos definir uma função (`regrOrigem`) que receba um ângulo ß e retorne uma visualização dos dados e o erro quadrático médio.

1. Primeiro, vamos criar variáveis `x` e `y` para receber os pontos centrais.
2. Vamos criar um novo data frame com esses pontos
   `dadosFr <- as.data.frame(table(x,y))`
3. Para efeitos de apresentação, vamos colocar os nomes novamente `'filhos', 'pais', 'frequência'.`

Hands On!

- Vamos definir uma função (`regrOrigem`) que receba um ângulo ß e retorne uma visualização dos dados e o erro quadrático médio.

4. Vamos criar um *plot* passando pais e filhos (note que a classe não é numérica!)

5. Como terceiro argumento, vamos alterar o tamanho dos pontos de acordo com sua frequência. `cex = fator_de_multiplocacao * frequencia`

Hands On!

- Vamos definir uma função (`regrOrigem`) que receba um ângulo ß e retorne uma visualização dos dados e o erro quadrático médio.

6. Fora do plot, vamos usar `abline` partindo de 0 até beta. Vamos usar pontos com 0,0 e pch=19 ou 20
7. O erro médio quadrático é calculado com mean(*y-beta*x)^2*.
8. Finalmente, usamos `title` para ver beta e o MSE.

* Chame *manipulate* com `regrOrigem(beta), beta=slider(0.5,1, step=0.01`

Hands On!

- Use a função criada para fazer um modelo para o conjunto de dados `father.son`
- Use para x e y: $\ldots = serie - media(serie)$
- Use `cor(x,y)` para obter a correlação
- Use internamente na sua função de plot:

```
plot(x, y,
     xlab = "Father's height, normalized",
     ylab = "Son's height, normalized",
     xlim = c(-3, 3), ylim = c(-3, 3),
     bg = "lightblue", col = "black", cex = 1.1, pch = 21,
     frame = FALSE)
```

- Crie duas linhas onde pai é preditor de filho e vice-versa.