Introdução à data science com R

Cleuton Sampaio

Sessão 1 – Visão geral

O que esperar?

- Conceitos;
- Introdução à estatística;
- Introdução à linguagem R;
- Regressão;
- Classificação.

Material e exemplos

- O material do curso está no Github:
  - https://github.com/cleuton/datascience/tree/master/R-course
  - Cada aula tem uma pasta "lessionX".

https://github.com/cleuton/datascience
Depois procure a pasta R-course

Uma pergunta interessante

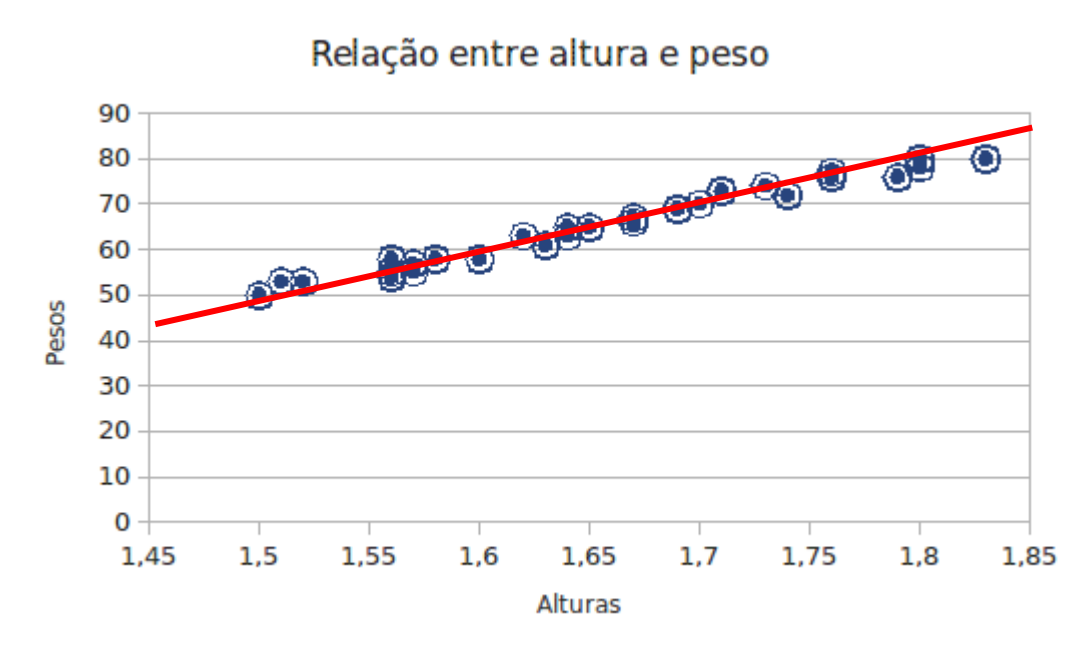
- Em uma determinada região, as autoridades estão preocupadas com o peso das crianças, com relação às suas alturas. Podemos estabelecer uma relação e tentar prever o peso dos estudantes com relação às suas alturas?

Coletamos amostras

- No material você encontrará o arquivo:
  - lesson1/mod-preditivo.ods
- Vamos abrir e examinar esta planilha. Se você não tiver o libreoffice instalado, pode usar a versão PDF, mas não terá como executar.

Reta de regressão

$$y = ax + b$$



- Queremos um modelo que, dada a altura, retorne o peso esperado;
- Uma reta cujas distâncias para os pontos reais seja a menor possível.

Solução de forma fechada

$$a = \frac{\sum_{i=0}^{n} (x_i - \bar{x}) \times \sum_{i=0}^{n} (y_i - \bar{y})}{\sum_{i=0}^{n} (x_i - \bar{x})^2}$$

$$b = \bar{y} - a \times \bar{x}$$

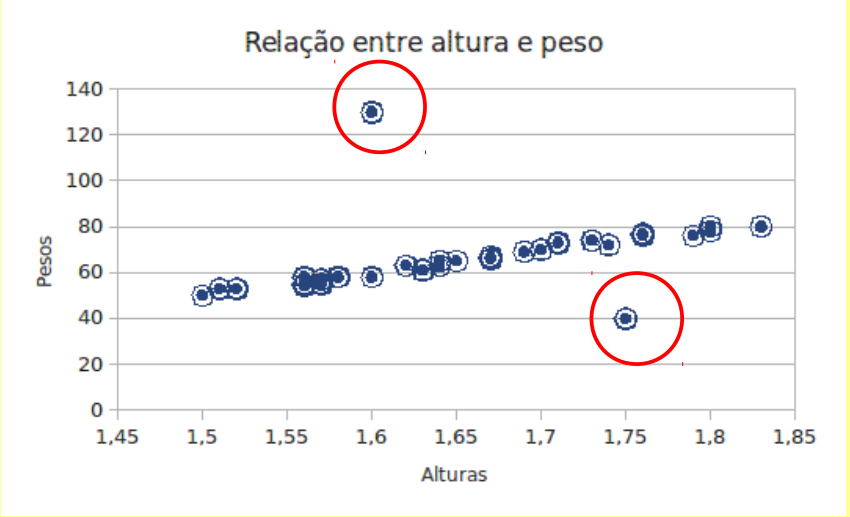
- Neste caso, podemos usar fórmulas para calcular os coeficientes da reta de regressão;
- Em outros casos, temos que usar heurísticas, como o método dos "Mínimos quadrados".

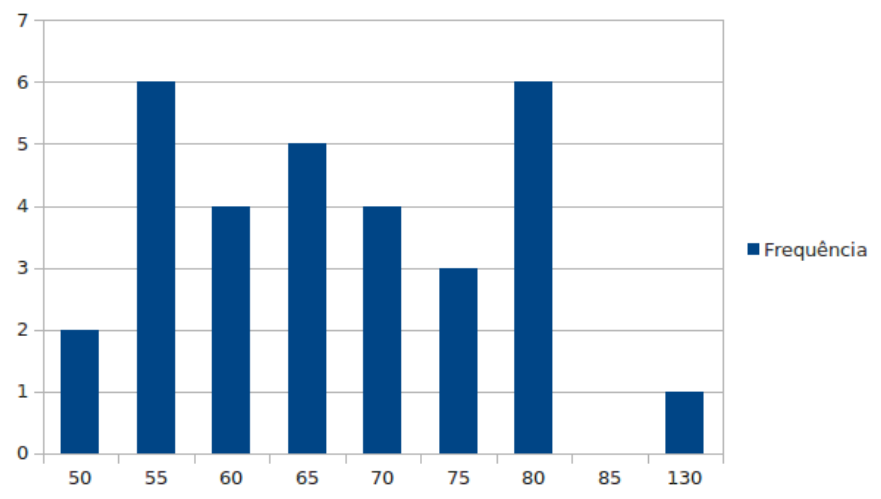
Antes de chegar nisso...

- Analisar os dados;
- Preparar os dados;
- Montar o modelo.

Analisando os dados

Pesos	Alturas	Pesos	Alturas
58	1,58	58	1,6
78	1,8	53	1,52
70	1,7	55	1,57
80	1,8	57	1,57
77	1,76	66	1,67
74	1,73	65	1,64
61	1,63	50	1,5
65	1,65	63	1,64
55	1,56	58	1,56
76	1,79	55	1,56
54	1,56	63	1,62
53	1,51	73	1,71
69	1,69	80	1,83
67	1,67	76	1,76
72	1,74	40	1,75
		130	1,6





Preparando os dados

- Desvio padrão:
  - Pesos: 15,52
  - Alturas: 0,09
- Limite de 3 desvios padrões:
  - Peso: 50 kg, altura: 1,50 m; → É possível
  - Peso: 40 kg, altura: 1,75 m; → Muito suspeito
  - Peso: 130 kg, altura: 1,6 m. → Muito suspeito

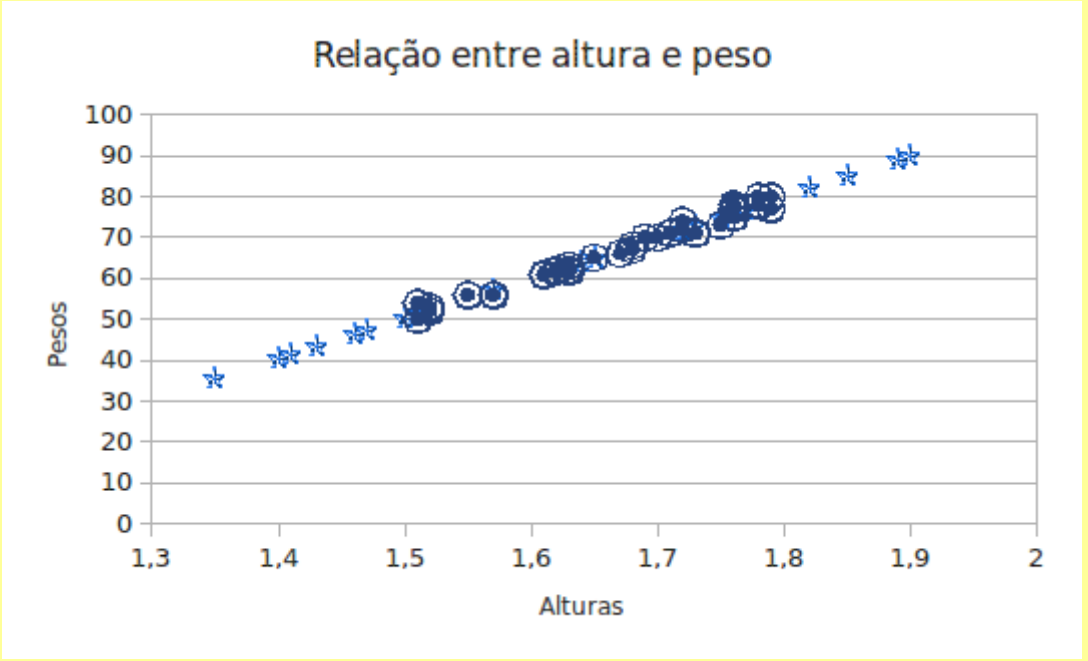
Mas isto é tudo?

- Regressão: Predizer um valor contínuo;
  - Estimar o preço de um imóvel, com base em suas características;
- Classificação: Predizer um valor discreto ou categório;
  - Classificar uma imagem (pessoa, carro, animal etc);
- Agrupamento: Separar os dados em grupos.
  - Identificar grupos separados de clientes, de acordo com os hábitos de compra.

Resultado

'y = ax + b	
a	95,7863963094089
b	-93,4166934878235

Nova altura	Novo Peso
1,46	46,4314451239137
1,65	64,6308604227014
1,51	51,2207649393842
1,57	56,9679487179487
1,64	63,6729964596073
1,72	71,33590816436
1,75	74,2095000536423



- Criamos um modelo preditivo para os pesos dos alunos

Como seria isso em R

- Vamos usar a linguagem R;
- Vamos usar o RStudio;
- Vamos ler a planilha e criar um modelo de regressão linear.

Não se preocupe! Você não vai entender muita coisa agora, mas terá uma visão geral do que vamos aprender. Relaxe e observe!
https://github.com/cleuton/datascience/tree/master/R-course/lession1
lerOds.R