Análise de agrupamento

Patrícia de Siqueira Ramos

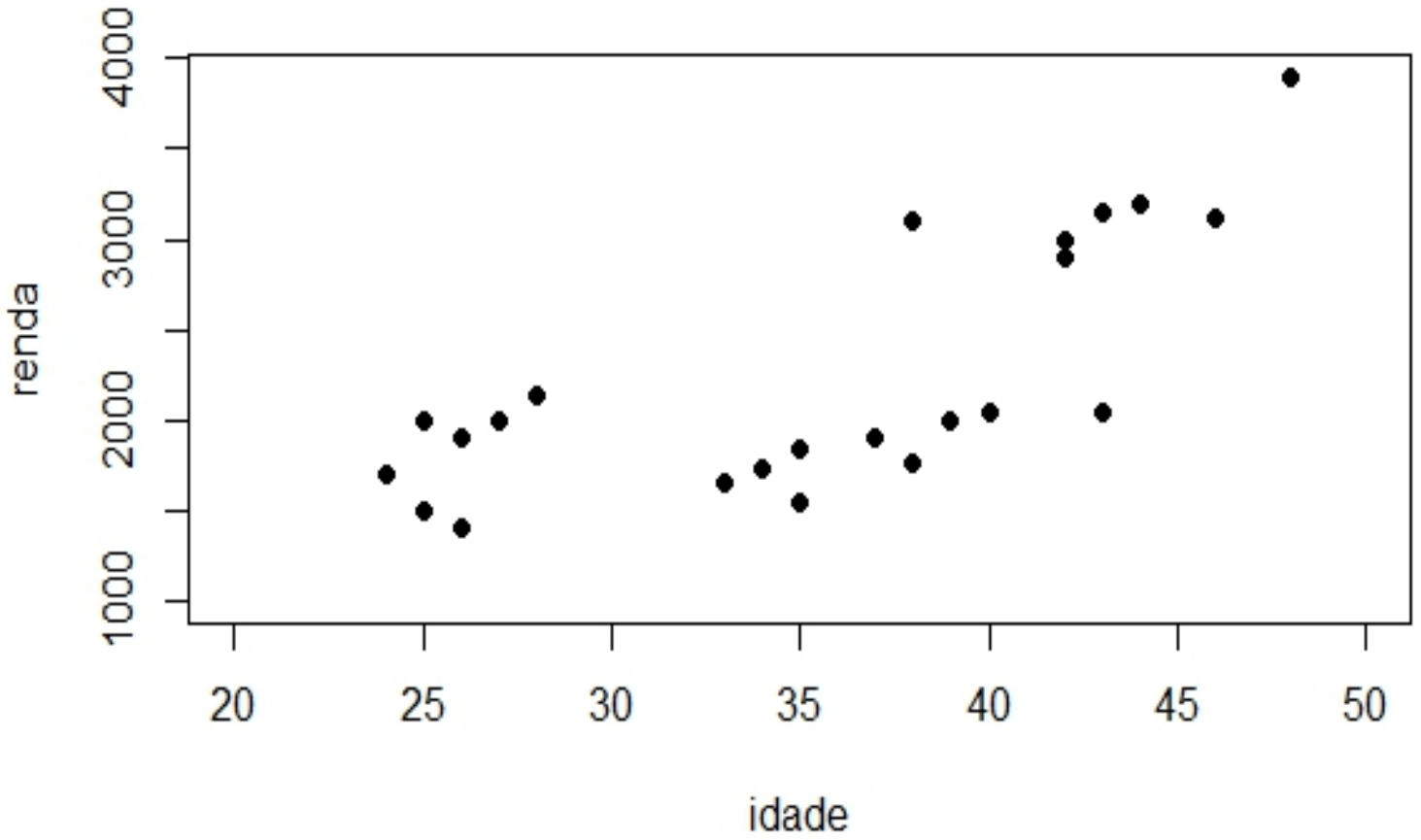
UNIFAL-MG, *campus* Varginha

18 de Junho de 2018

Métodos de agrupamento

- Análise de agrupamento = conglomerados, classificação, *cluster analysis*
- Objetivo: agrupar elementos da amostra de forma que:
  (i) observações de um mesmo grupo sejam similares entre si em relação às variáveis medidas
  (ii) observações de grupos deferentes sejam heterogêneas em relação a essas variáveis
- AA tenta formalizar o que os seres humanos fazem bem em duas ou três dimensões (diagramas de dispersão)
- Os grupos são identificados com base nas distâncias entre os pontos

Ex.: Dados de idade e renda de 23 pessoas



Exemplo de renda e idade

- Se fôssemos agrupar as pessoas em relação às variáveis:
  - pela análise visual do diagrama: quantos grupos teríamos?
  - e só avaliando a renda, quantos grupos seriam?

Objetivos possíveis de um agrupamento

1. agrupar as n observações em grupos não definidos antes (AA - análise de agrupamento)
2. classificar as n observações em um conjunto predefinido de grupos (AD - análise discriminante)

Objetivos possíveis de um agrupamento

1. agrupar as n observações em grupos não definidos antes (AA - análise de agrupamento)
2. classificar as n observações em um conjunto predefinido de grupos (AD - análise discriminante)

- Em AA, encontrar o melhor agrupamento não é simples
- Cada observação tem informações de p variáveis em um vetor e, para compará-las, usamos métricas para comparar vetores, como distâncias

Distância

- Ao efetuar um agrupamento de observações, é preciso saber se as observações estão próximas ou distantes
- Se há n observações e p variáveis, cada observação i é representada por

$$\mathbf{x}_{i\cdot} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix},$$

em que X_{ij} é o valor observado da variável j no elemento i

Matriz de distâncias **D**

Inicia-se com uma matriz de distâncias **D** de dimensão $n \times n$: dois elementos i e j são considerados próximos quando sua distância é pequena (ou a similaridade é grande/ dissimilaridade é pequena):

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & n_{n2} & \dots & d_{nn} \end{bmatrix},$$

(simétrica) em que d_{ij} é a distância entre as observações i e j

Distâncias

Antes do agrupamento em si é preciso decidir qual medida de similaridade será usada (há várias e que produzem diferentes grupos):

- Tipo mais simples e comum: distância euclidiana:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2} \qquad \text{ou} \qquad d_{ij}^2 = (\mathbf{X_{i.}} - \mathbf{X_{j.}})^T(\mathbf{X_{i.}} - \mathbf{X_{j.}})$$

- distância generalizada:

$$d_{ij}^2 = (\mathbf{X_{i.}} - \mathbf{X_{j.}})^T\mathbf{A}(\mathbf{X_{i.}} - \mathbf{X_{j.}}),$$

em que $\mathbf{A}$ é uma matriz P.D. e pode ser:

- $\mathbf{A} = \mathbf{I}$: distância euclidiana
- $\mathbf{A} = \mathbf{D}^{-1}$: distância euclidiana padronizada ($\mathbf{D}^{-1} = \text{diag}(1/S_{ii})$)
- $\mathbf{A} = \mathbf{S}^{-1}$: distância de Mahalanobis

Distâncias

A escolha de **A** depende do tipo de informação que o pesquisador quer levar em conta ao comparar as observações:

- variáveis em escalas similares: euclidiana
- variáveis em diferentes escalas mas pouco correlacionadas: euclidiana padronizada
- variáveis em diferentes escalas e diferentes estruturas de correlações: Mahalanobis

Distâncias

Há ainda:

- distância de Minkowski:

$$d_{ij} = \left(\sum_{k=1}^{p} |X_{ik} - X_{jk}|^{\alpha} \right)^{1/\alpha}.$$

- se $\alpha = 1$: distância *city-block* (quarteirão) ou Manhattan (pouco influenciada por *outliers*)
- se $\alpha = 2$: distância euclidiana

Exercício

Dados de renda mensal e idade de seis pessoas:

indivíduo	renda	idade
1	9,60	28
2	8,40	31
3	2,40	42
4	18,20	38
5	3,90	25
6	6,40	41
$\bar{X}$	8,15	34,17
S	5,61	7,14

Obter a distância entre as observações 1 e 2 usando todas as métricas vistas.