Introdução à Análise de Dados Utilizando o Ambiente R

Marcelo de Souza Lauretto
Sistemas de Informação – EACH
marcelolauretto@usp.br

Curso de Verão EACH/USP
Fevereiro / 2015

Agenda

- R: Definição e história
- R Commander:
  - Breve tutorial
  - Análise exploratória de dados
  - Testes de hipóteses
  - Análise de agrupamentos
  - Regressão Linear

Referências

- J. Fox. Using the R Commander: A Point-and-Click Interface for R. Chapman&Hall / CRC Press, 2017. http://socserv.mcmaster.ca/jfox/Books/RCommander/
- Torsten Hothorn and Brian S. Everitt. A Handbook of Statistical Analyses Using R. Chapman & Hall/CRC Press, Boca Raton, Florida, USA, 3rd edition, 2014. http://www.crcpress.com/product/isbn/9781482204582
- W. N. Venables, D.M.Smith and the R Core Team. An Introduction to R. Version 3.1.2, 2014. http://www.cran.r-project.org/doc/manuals/R-intro.pdf

Referências

- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. http://www.R-project.org
- R. Ihaka. R: Past and Future History. Statistics Department, The University of Auckland, Auckland, New Zealand. http://cran.r-project.org/doc/html/interface98-paper/paper.html
- J.Fox, M. Bouchet-Valat. Getting Started With the R Commander. Version 2.3-0. http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf

R: Conceitos Básicos

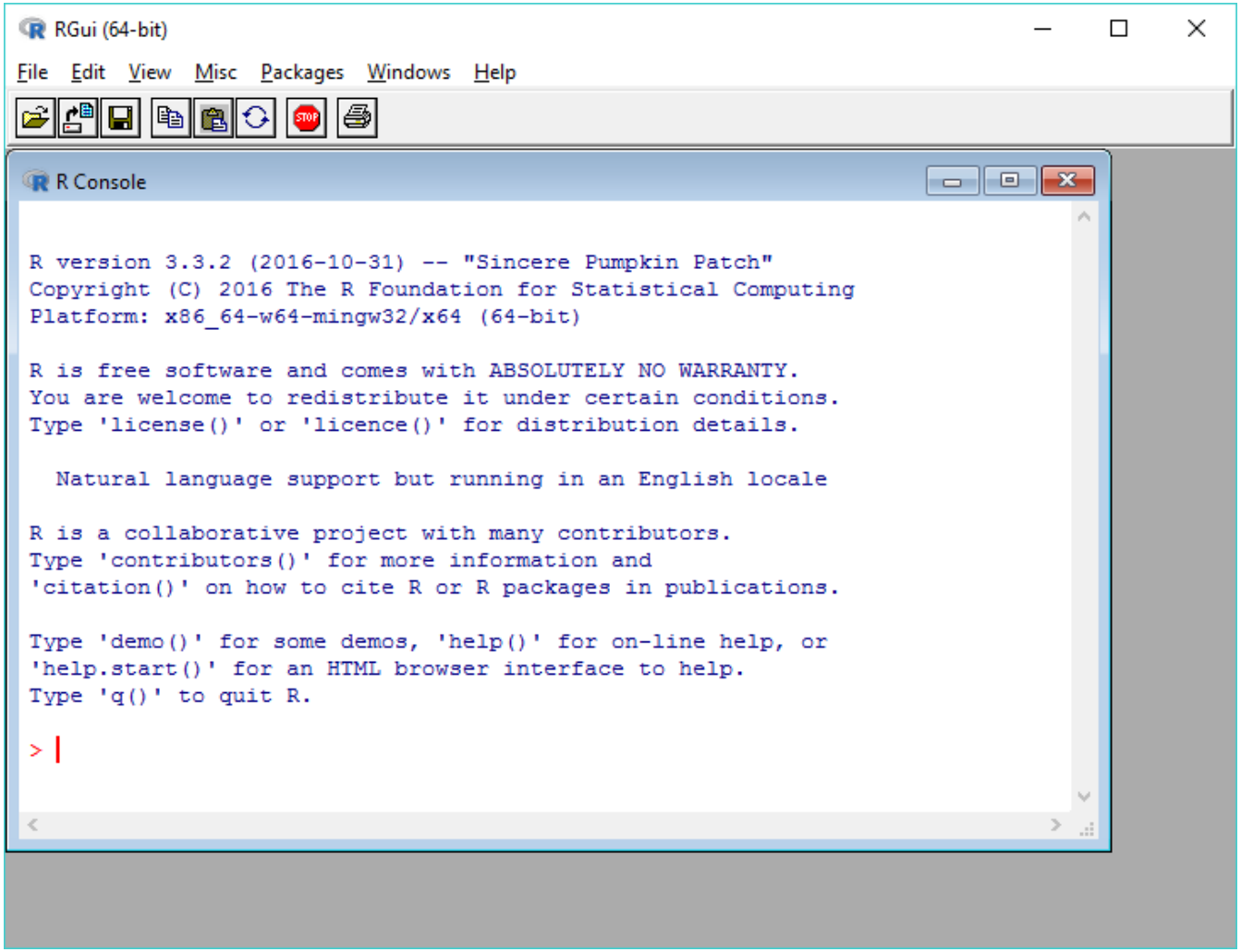
R: definição e história

- R é uma linguagem e um ambiente de desenvolvimento voltado principalmente para computação estatística (inferência, simulações, *data mining*, etc) e gráficos.
- Inspirado em duas linguagens:
  - S (John Chambers e colegas – Bell Labs): sintaxe
  - Scheme (Hal Abelson and Gerald Sussman): implementação e semântica
- Desenvolvido originalmente por Ross Ihaka e Robert Gentleman (Depto Estatística da Universidade de Auckland, Nova Zelândia).
- Atualmente desenvolvido pelo *R Development Core Team*
- R está disponível como um software livre, nos termos da GNU GPL (General Public License).
  - Windows, Linux, OS X (Mac)

O Projeto R

- Software e documentação disponível em www.r-project.org
- Conteúdo geral do site:
  - Sobre o R
  - Download, packages:
    - CRAN (Comprehensive R Archive Network)
  - Documentação
    - Manuals
    - FAQs (Frequently Asked Questions)
  - Informações suplementares:
    - CRAN Task views: guias para pacotes e funções úteis para certas áreas/disciplinas
    - Ferramenta de busca no site (opção *Search*), muito útil

Console do R


RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
STOP

R: Ajuda

- Ajuda no ambiente R:

  help.start()

  help('while'), help('print'), ?print

  help.search('regression')

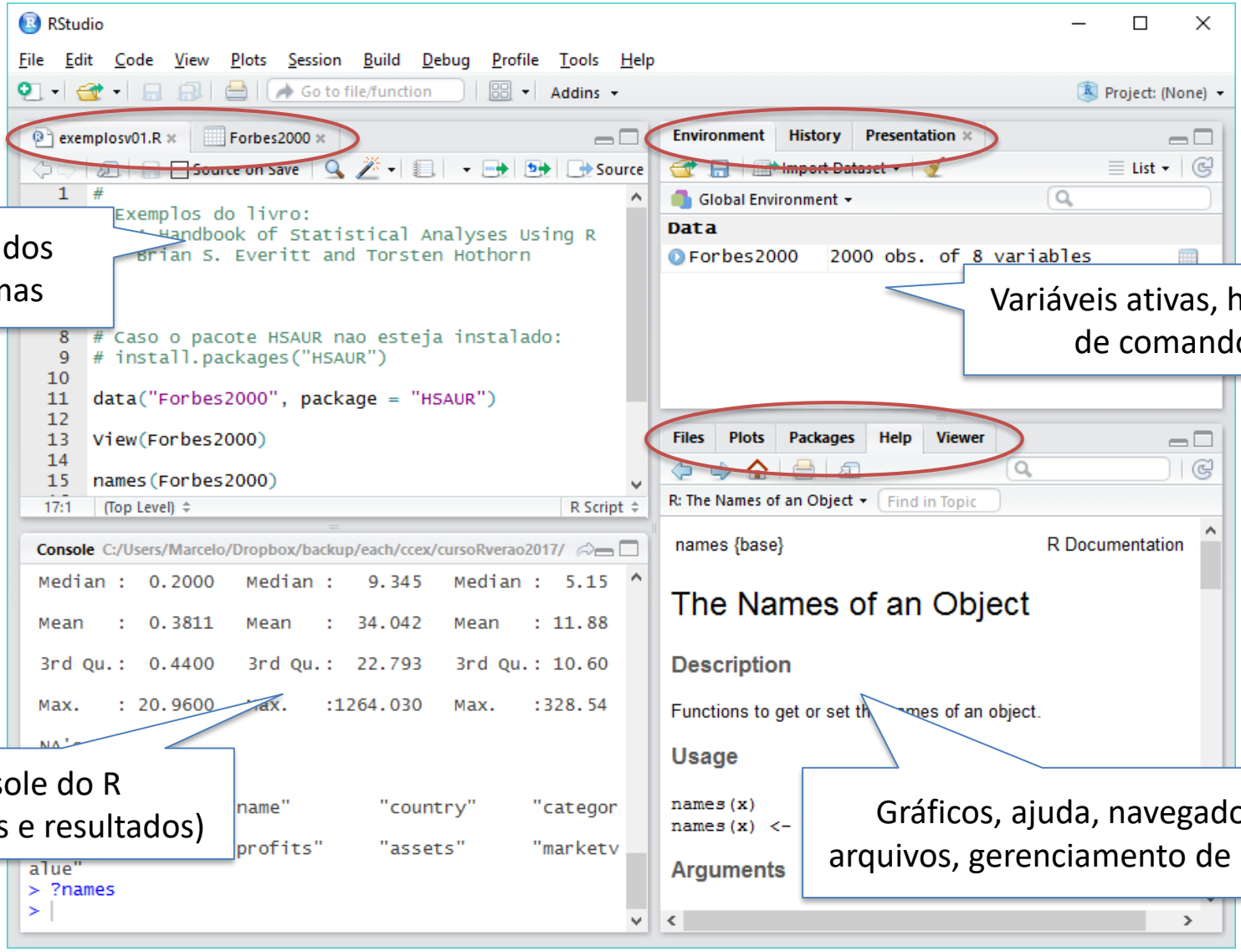
- R site search:
  - http://finzi.psych.upenn.edu/search.html
- CRAN Task Views:
  - http://cran.r-project.org/web/views/
- Cartão de referência preparado por Jonathan Baron:
  - http://www.leg.ufpr.br/~paulojus/misc/refcard.pdf

RStudio: Ambiente de desenvolvimento

RStudio

- RStudio é um ambiente de desenvolvimento integrado (IDE – Integrated Development Environment) para R
- Integração de:
  - Editor de programas
  - Facilidade de execução parcial ou total de scripts R
  - Visualização de dados (Tabelas, Gráficos)
  - Documentação (Help)
  - Ferramentas de depuração de programas
- Disponível para Windows, Linux, OS-X (Mac)
- Website oficial:
  - https://www.rstudio.com

Ambiente do RStudio



Códigos dos programas

Variáveis ativas, histórico de comandos

Console do R (comandos e resultados)

Gráficos, ajuda, navegador de arquivos, gerenciamento de pacotes

Linguagem R

Referência:

P. J. Ribeiro Jr. Introdução ao Ambiente Estatístico R. 2011.

http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa/

Tipos de dados no R

- Tipos de objetos mais usuais em R:
  - vector: o mais elementar e um dos mais importantes
  - matrix e array: generalizações multi-dimensionais de vetores
  - data frame: conjunto de dados retangular no qual:
    - linhas representam os casos (sujeitos do estudo)
    - colunas representam as variáveis descritivas dos casos
- Cada objeto em R (e cada coluna em um data frame) possui um dos seguintes tipos básicos:
  - numeric: para variáveis numéricas (reais, complexas ou inteiras)
  - factor: representação de variáveis categóricas nominais ou ordinais
  - logical: TRUE/FALSE
  - character: texto (string)

Vetores

- Atribuição:

  x = c(10.4, 5.6, 3.1, 6.4, 21.7)

  Operador de atribuição: = , <- ou ->

  x

  mode (x)

  length(x)

  y = c(x, 0, x)

- Aritmética:

  v = 2*x + y + 1

- Funções estatísticas e sumários:

```
sum(x)
length(x)
mean(x)   #  equivalente a sum(x) / length(x)
var(x)       # equivalente a sum((x-mean(x))^2) / (length(x)-1)
```

- Sequências regulares

```
s1 = 1:30
n=10
s2 = 1:n-1
s3 = 1:(n-1)
s4 = seq(-5, 5, by=.2)
```

- Vetores lógicos

  temp = x>13

- Vetores de índices e Filtros

  idx123 = 1:3

  x[idx123]

  idxval = which(x>13)

  x[idxval]

- Vetores de caracteres

  letras = c('a', 'b', 'c')

  repeticao_a

  labs <- paste(c("X","Y"), 1:10, sep='')

Operações e funções matemáticas:

```
2 + 4 * 5      # Order of operations
log (10)       # Natural logarithm with base e=2.7182
log10(5)       # Common logarithm with base 10
5^2            # 5 raised to the second power
5/8            # Division
sqrt (16)      # Square root
abs (3-7)      # Absolute value
pi             # 3.14
exp(2)         # Exponential function
round(pi,0)  # Round pi to a whole number
round(pi,1)  # Round pi to 1 decimal place
round(pi,4)  # Round pi to 4 decimal places
```

Operações e funções matemáticas:

```
floor(15.9)   # Rounds down
ceiling(15.1)  # Rounds up
cos(.5)       # Cosine Function
sin(.5)       # Sine Function
tan(.5)       # Tangent Function
acos(0.8775826)       # Inverse Cosine
asin(0.4794255)       # Inverse Sine
atan(0.5463025)      # Inverse Tangent
```

Outros tipos de objetos

- Matrizes:

```
M = matrix(1:20, nrow=5, ncol=4)
N = matrix(1:20, nrow=5, ncol=4, byrow=TRUE)
colnames(N) = c('a', 'b', 'c', 'd')
N[, 1:2]
N[c(2,4),]
N[c(2,4), 1:2]    # acesso a porcoes especificas
N[,c('a', 'b')]
```

Outros tipos de objetos

- Data-frames:
  - Vetores e matrizes forçam que todos os elementos sejam do mesmo tipo
  - Data-frames não possuem essa restrição. Cada coluna pode ser de um tipo diferente.

```
d1 = data.frame(X=1:10, Y=c(rep('low',5), rep('upp',5)))
summary(d1)
d1$X              # exibe o conteudo da 1a coluna
d1[ ,c('X','Y')]  # forma util de selecionar multiplas colunas
```