Bài giảng 1:
Ôn tập căn bản về R

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

University of Technology Sydney (UTS) and UNSW Australia

Ton Duc Thang University, Vietnam

Ngôn ngữ (phần mềm) R

- **Mã nguồn mở** - hoàn toàn miễn phí !
- Chạy trên Windows, Unix, MacOS.
- Do các chuyên gia thống kê phát triển
- Rất nhiều phương pháp phân tích, cơ bản đến nâng cao
- Biểu đồ chất lượng cao
- Các đại học và viện nghiên cứu rất chuộng R

Căn bản R

- Vận hành
- Đọc dữ liệu vào R
- Biến số trong R
- Biên tập số liệu

Vận hành R

Object = đối tượng

- R vận hành theo đối tượng: biến số, dataset, input, output, v.v. đều được xem là *object*
- Đối tượng phải có *tên*
- Tên: phân biệt chữ hoa và chữ thường

tuan, TUAN, Tuan khác nhau

Packages

- **R = Base + Packages**
- **Base** là phần mềm cơ bản bao gồm một số hàm dùng cho phân tích dữ liệu
- **Packages** là những modules dùng cho các phân tích chuyên dụng
- Có hơn 6000 packages trong R
- Có thể tải về và cài đặt packages trực tiếp từ mạng

Một số packages phổ biến

Hmisc: Miscellaneous for data manipulation

tables: For tabulation of data

foreign: For reading data from other softwares

gmodels: Programming tools

ggplot2: Advanced graphics

sciplot: Scientific graphs

Zelig: "Every one's statistical software"

MASS: Phương pháp thống kê hiện đại

rms: Regression modeling strategies

car: Companion to regression analysis

survival: Survival analyses

EpiR: Epidemiological analyses

epicalc: Epidemiological analyses

boot: Bootstrap analyses

cluster: Cluster analysis

psych: Psychometrics and descriptive statistics

Cài đặt package (ví dụ)

```
install.packages(c("Hmisc", "rms", "tables",
"foreign", "gmodels", "ggplot2", "sciplot",
"Zelig", "car", "survival", "EpiR",
"epicalc", "boot", "cluster", "psych",
"binom", "BMA", "ExactCIdiff", "lattice",
"mgcv", "gam", "nlme", "quantreg")
```

- Tìm xem đang có package nào đã được cài đặt

```
library()
```

Văn phạm R

Tương tác với R

- Khởi động R
- Dùng mũi tên up/down arrow keys để tìm các lệnh trước trong console
- Dùng mũi tên left/right keys để chỉnh sửa (edit) lệnh
- Dùng TAB để có thêm lựa chọn (rất có ích)
- Có thể viết nhiều lệnh trong 1 dòng, cách nhau bằng dấu ";"

Tên biến số (variable)

- Dùng mẫu tự, số, kí hiệu (., -, _)
- Kí hiệu "assignment": <- hoặc =
- Phân biệt chữ thường và chữ hoa

```
Genotype = 5; genotype <- 7;

Geno.type = Genotype + genotype
```

Hàm (function)

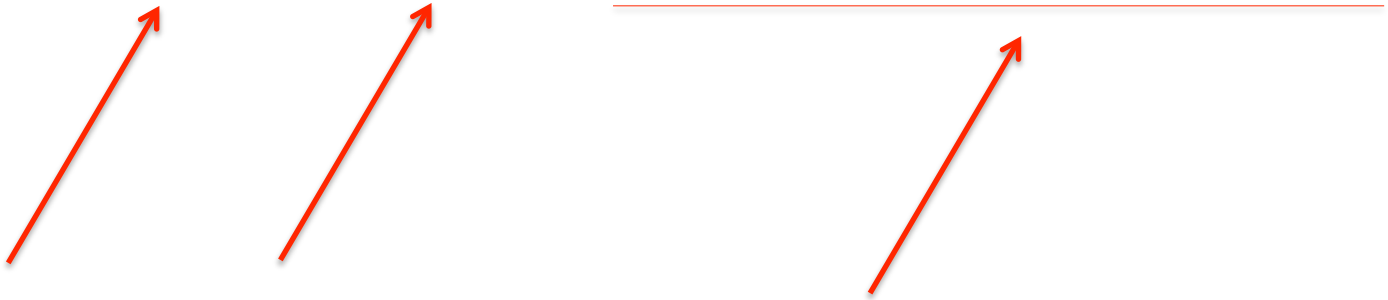
- Lệnh R = function
- Hàm phải có arguments
- Arguments bao gồm variable, parameters, options, v.v.
- Ví dụ: Phân tích mô hình y = a + bx

```
m1 = lm(y ~ x, data=test)
```

Function

- Lệnh R = function
- Hàm phải có arguments
- Ví dụ: Phân tích mô hình y = a + bx

```
m1 = lm(y ~ x, data=test)
```



Object name
m1

Function
lm = linear model

Arguments:
variables: y, x
dataset name

Đọc dữ liệu

Các dữ liệu R có thể đọc

- Đọc trực tiếp
- ASCII và text files
- Excel / **csv**
- SAS, SPSS, Stata, etc.
- Databases

Đọc dữ liệu trực tiế: c()

age	sex	weight
18	M	60.3
21	F	48.5
35	M	62.0
50	F	47.2

```
age = c(18, 21, 35, 50)
sex = c("M", "F", "M", "F")
weight = c(60.3, 48.5, 62.0, 47.2)

# tạo thành dataset

dat = data.frame(age, sex, weight)
dat
```

```
> dat
  age sex weight
1  18   M   60.3
2  21   F   48.5
3  35   M   62.0
4  50   F   47.2
```

Đọc từ ascii files: read.table

File: "Hoa hau.txt"

YoB	Year	Height	Weight
1971	1988	157	50
1969	1990	158	NA
1976	1992	174	NA
1976	1994	172	NA
1976	1996	170	NA
1976	1998	172	50
1980	2000	169	49
1985	2002	169	49
1985	2004	172	52
1988	2006	181	60
1990	2008	182	61.5
1989	2010	173	55
1991	2012	173	49
1996	2014	173	59

Đọc từ ascii files: read.table

```
hh = read.table("~/Dropbox/hoa hau.txt",
header=T, na.strings="NA")

hh
```

Giải thích

read.table() – hàm R

header = dùng dòng đầu trong file làm tên của biến số

na.strings = "NA", lấy NA làm kí hiệu cho giá trị khống (missing values)

```
> hh
    YoB Year Height Weight
1  1971 1988    157   50.0
2  1969 1990    158     NA
3  1976 1992    174     NA
4  1976 1994    172     NA
5  1976 1996    170     NA
6  1976 1998    172   50.0
7  1980 2000    169   49.0
8  1985 2002    169   49.0
9  1985 2004    172   52.0
10 1988 2006    181   60.0
11 1990 2008    182   61.5
12 1989 2010    173   55.0
13 1991 2012    173   49.0
14 1996 2014    173   59.0
```

Đọc dữ liệu từ excel

- Phức tạp (do cấu trúc excel thay đổi theo phiên bản)
- Cách tốt nhất:
  - "Xuất khẩu" sang dạng csv
  - Dùng hàm **read.csv()**

ID	Province	Subregion	Region	Year2014	Year2012	Year2011	Year2010	Year2009	Year2008	Year2007	Year2006
1	An Giang	DBSCL	Nam	99.64		90.3	81	75.2	79.9	71.7	77.8
2	Ba ria - Vung tau	Dong Nam Bo	Nam	99.46		97.21	92.58	84.57	70.2	69.9	93.4
3	Bac Giang	Dong Bac	Bac	99.47	99.04	99.37	97.8	88.04	82.3	60.6	97.6
4	Bac Kan	Dong Bac	Bac	98.98		88.7	70	60.95	43.2	20.3	91.2
5	Bac Lieu	DBSCL	Nam			96	85.35	73.08	65.2	48.8	79.2
6	Bac Ninh	DBSH	Bac			99.62	99.28	94.15	87	75.3	99.6
7	Ben Tre	DBSCL	Nam	99.67		84.15	72.29	79.71	83.3	79.8	86.4
8	Binh Dinh	Nam Trung Bo	Trung	99.17	99.6	96.84	93.9	88.48	82.2	71.4	95.4
9	Binh Duong	Dong Nam Bo	Nam	99.86		90.7	87.75	77.89	63.5	62.5	87.5
10	Binh Phuoc	Dong Nam Bo	Nam	99.41	99.56	94.57	92.04	82.19	67.2	58.4	93.4
11	Binh Thuan	Dong Nam Bo	Nam	98.48		88.06	83.2	81.73	75.6	73.6	89.5
12	Ca Mau	DBSCL	Nam	98.27	99.02	93.16	90.01	82.25	72.7	63.5	82.4
13	Can Tho	DBSCL	Nam	99.72	99.68	97.74	86	77.42	86.4	79.6	94.9
14	Cao Bang	Dong Bac	Bac	99		93.73	89.65	64.24	40.6	27.8	86.8
15	Da Nang	Nam Trung Bo	Trung	98.54	99.53	97.2	96.68	89.74	83.2	76.3	97.5
16	Dak Lak	Tay Nguyen	Trung	97.98	97.46	88.36	78.11	69.11	55.6	51	86.3
17	Dak Nong	Tay Nguyen	Trung	97.93		81.95	78.2	76.09	62.3	50.5	79.3
18	Dien Bien	Tay Bac	Bac	98.11		95.65	71	73.32	74.8	46	81.8

```
tn = read.csv("~/Dropbox/THPT.CSV", header=T,
na.strings=" ")

tn
```

Đọc dữ liệu từ stata

- File: truonghoc.dta
- Để đọc vào R chúng ta cần
  - Package "**`foreign`**"
  - Dùng hàm **`read.dta()`**

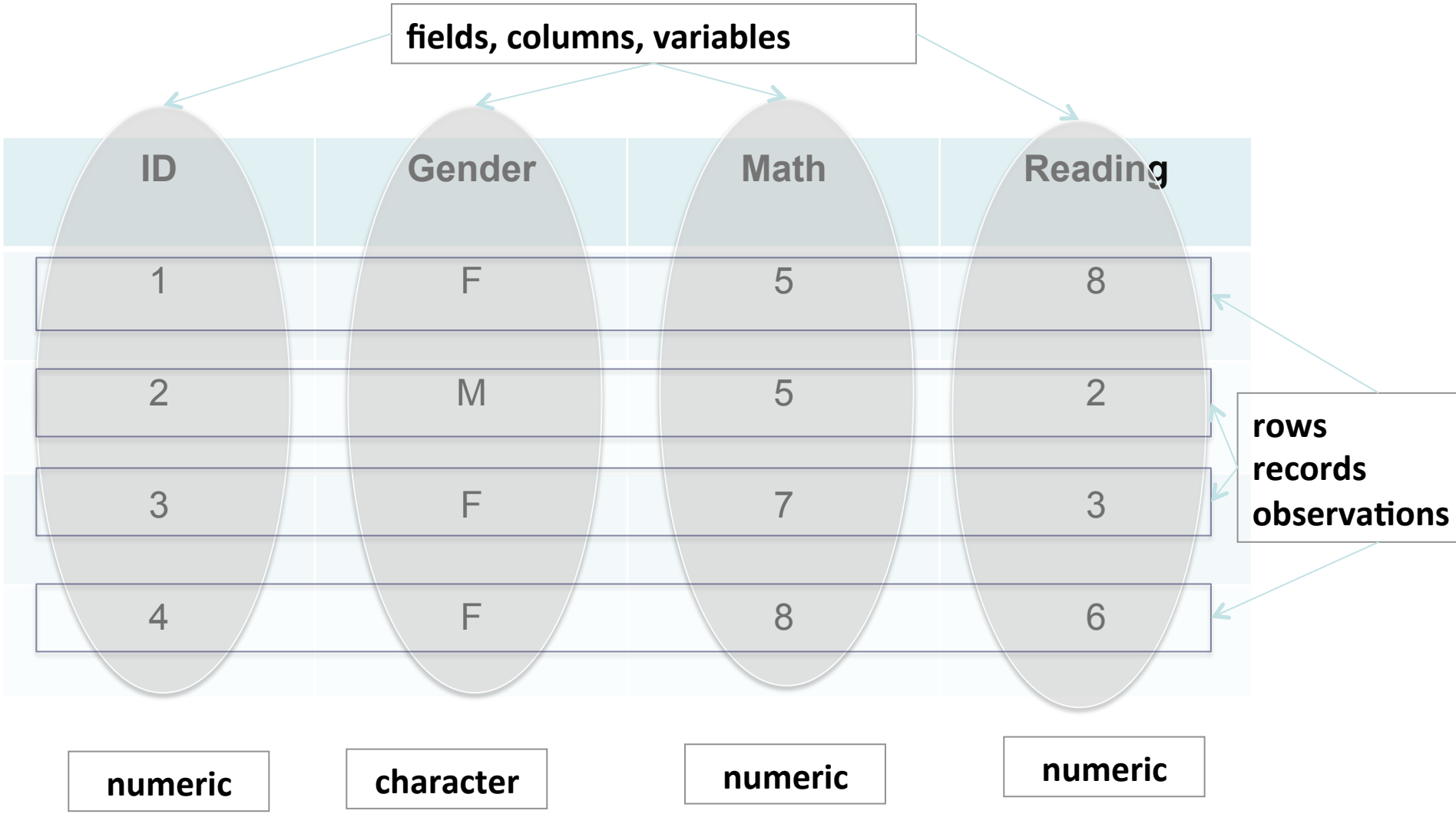
Đọc dữ liệu từ stata

```
library(foreign)

schools = read.dta("~/Dropbox/truonghoc.dta")

head(schools)
```

Làm việc với biến số

Dataframe

Dataset trong R = “Dataframe” = matrix



Đề cập đến biến

- Dataframe cần phải **`attach`** trước khi phân tích
- Đề cập chuẩn:

(dataframe name)**$**(field name)

- Ví dụ:

```
v1 = c(1,3,5)
v2 = c(3,4,7)
v3 = c(6,7,8)
v4 = c(7,9,0)
dat = data.frame(v1, v2, v3, v4)
attach(dat)
dat$sum = dat$v1 + dat$v3
sum1 = v1 + v3
dat
```

Ảnh hưởng của $

```
v1 = c(1,3,5)
v2 = c(3,4,7)
v3 = c(6,7,8)
v4 = c(7,9,0)
dat=data.frame(v1,v2,v3,v4)
attach(dat)
dat$sum = dat$v1 + dat$v3
sum1 = v1 + v3
dat
```

```
> dat
  v1 v2 v3 v4 sum
1  1  3  6  7   7
2  3  4  7  9  10
3  5  7  8  0  13
```

Không có Sum1 !

Biến số (variables)

- Biến số = cột dữ liệu
- File: SCHOOL DATA (VN).CSV

```
schools = read.csv("~/Dropbox/World Bank
  2014/Data for 2015 workshop/SCHOOL DATA
  (VN).csv", header=T)

attach(schools)

head(schools)

dim(schools)
```

```
> head(schools)
   REGION   TYPE  AREA
1 CENTRAL PUBLIC URBAN
2   NORTH PUBLIC URBAN
3   SOUTH PUBLIC RURAL
4   SOUTH PUBLIC URBAN
5 CENTRAL PUBLIC URBAN
6   NORTH PUBLIC URBAN
                                                    STRATUM SCHOOLID SC01Q01
1  VNM - stratum 07 : Central Viet Nam / Public / Urban        1  Public
2    VNM - stratum 01 : North Viet Nam / Public / Urban        2  Public
3 VNM - stratum 14 : Southern Viet Nam / Public / Rural        3  Public
4 VNM - stratum 13 : Southern Viet Nam / Public / Urban        4  Public
5  VNM - stratum 07 : Central Viet Nam / Public / Urban        5  Public
6    VNM - stratum 01 : North Viet Nam / Public / Urban        6  Public
  SCHSIZE SC09Q11    SC03Q01     SC04Q01 SC05Q01 CLSIZE COMPWEB PCGIRLS
1    1804      93 Small Town   One Other     >50     53      NA   0.557
2    1586      84       Town Two or More   36-40     38       1   0.505
3     604      32    Village   No Others   41-45     43       0   0.533
4     568      99 Small Town   No Others   36-40     38       1   0.586
5    1078      65 Small Town   One Other   41-45     43       1   0.552
6    1232      37 Small Town Two or More   41-45     43       1   0.594
  SCMATEDU SMRATIO STRATIO
1  -1.0530 120.267  18.890
2  -0.5214 105.733  18.230
3  -1.9620 151.000  18.585
4  -1.2473  35.500   5.737
5   0.2240  98.000  16.211
6   0.0288      NA  33.297
```

Chúng ta phân tích theo biến

- table(REGION)
- table(REGION, AREA)
- mean(**SMRATIO)**
- mean(**SMRATIO, na.rm=T)**

Dữ liệu dạng tóm lược (summary data)

Region	Remote	Rural	Urban
Central	5	27	22
North	8	25	24
South	2	22	27

Region	Area	N
Central	Remote	5
Central	Rural	27
Central	Urban	22
North	Remote	8
North	Rural	25
North	Urban	24
South	Remote	2
South	Rural	22
South	Urban	27

Dữ liệu dạng tóm lược (summary data)

Region	Area	N
Central	Remote	5
Central	Rural	27
Central	Urban	22
North	Remote	8
North	Rural	25
North	Urban	24
South	Remote	2
South	Rural	22
South	Urban	27

```
Region = c(rep("Central",
  3), rep("North", 3),
  rep("South", 3))
Area = rep(c("Remote",
  "Rural", "Urban"), 3)
N = c(5, 27, 22, 8, 25, 24,
  2, 22, 27)

dat = data.frame(Region,
  Area, N)
data
```

Biên tập dữ liệu

Mã hoá (coding)

```
id = c(1, 2, 3, 4, 5)

gender = c("male", "female", "male", "female",
"female")

dat = data.frame(id, gender)
```

We want to create a new variable called `sex` with numeric values (1, 2)

```
dat$sex[gender=="male"] <- 1

dat$sex[gender=="female"] <- 2
```

Character và numeric

Từ character chuyển sang numeric

```
X = c("1", "2", "3", "4", "5")
```

We want to create a new variable called Y with numeric values (for calculation)

```
Y = as.numeric(X)
```

```
mean(Y)
```

Từ numeric chuyển sang character

```
Y = 1:10
```

We want to create a new variable called X with character values

```
X = as.character(Y)
```

Numeric và factor

FACTOR cần thiết cho phân tích và phân nhóm

Từ numeric chuyển sang factor

```
Y = 1:3
```

We want to create a new variable called X with character values

```
X = as.factor(Y)
```

sort()

```
X = rnorm(10); X
```

```
[1]  1.5651300 -0.5382971 -0.1995302  1.0111098  0.3590144 -1.5245237
[7] -0.3192534  0.1323256 -0.7916954 -0.0664167
```

```
sort(X)
```

```
[1] -1.5245237 -0.7916954 -0.5382971 -0.3192534 -0.1995302 -0.0664167
[7]  0.1323256  0.3590144  1.0111098  1.5651300
```

Hợp nhất dữ liệu: merge()

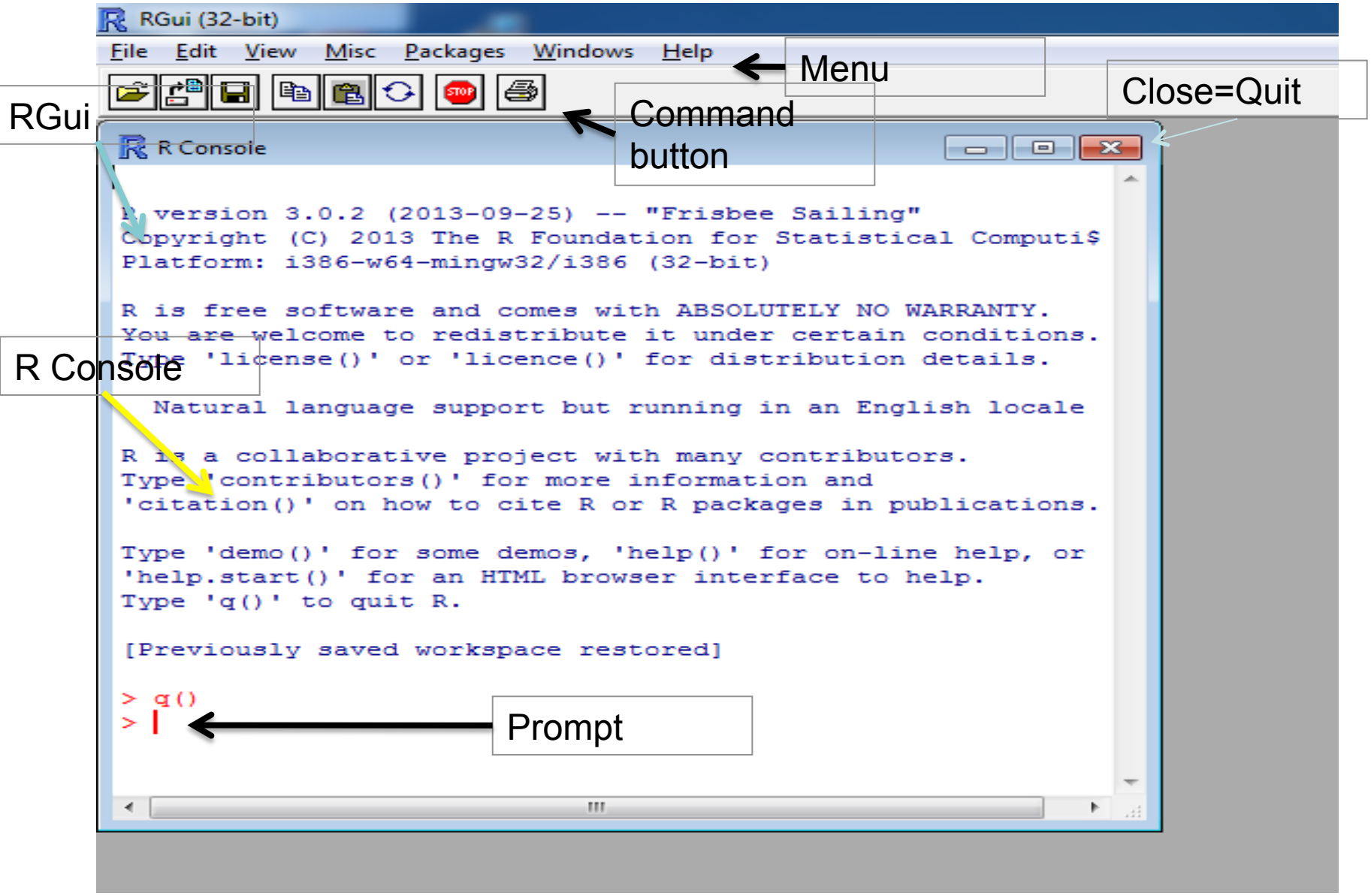
```
id = c(1,2,3,4)
sex=c("M","F","M","F")
dat1=data.frame(id,sex)
```

```
id = c(1,2,3,4,5)
age=c(21,34,45,32,18)
dat2=data.frame(id,age)
```

```
dat = merge(dat1, dat2, by="id")

dat = merge(dat1, dat2, by="id", all.x=T, all.y=T)
```

R và RStudio

Một phiên làm việc với R



RStudio

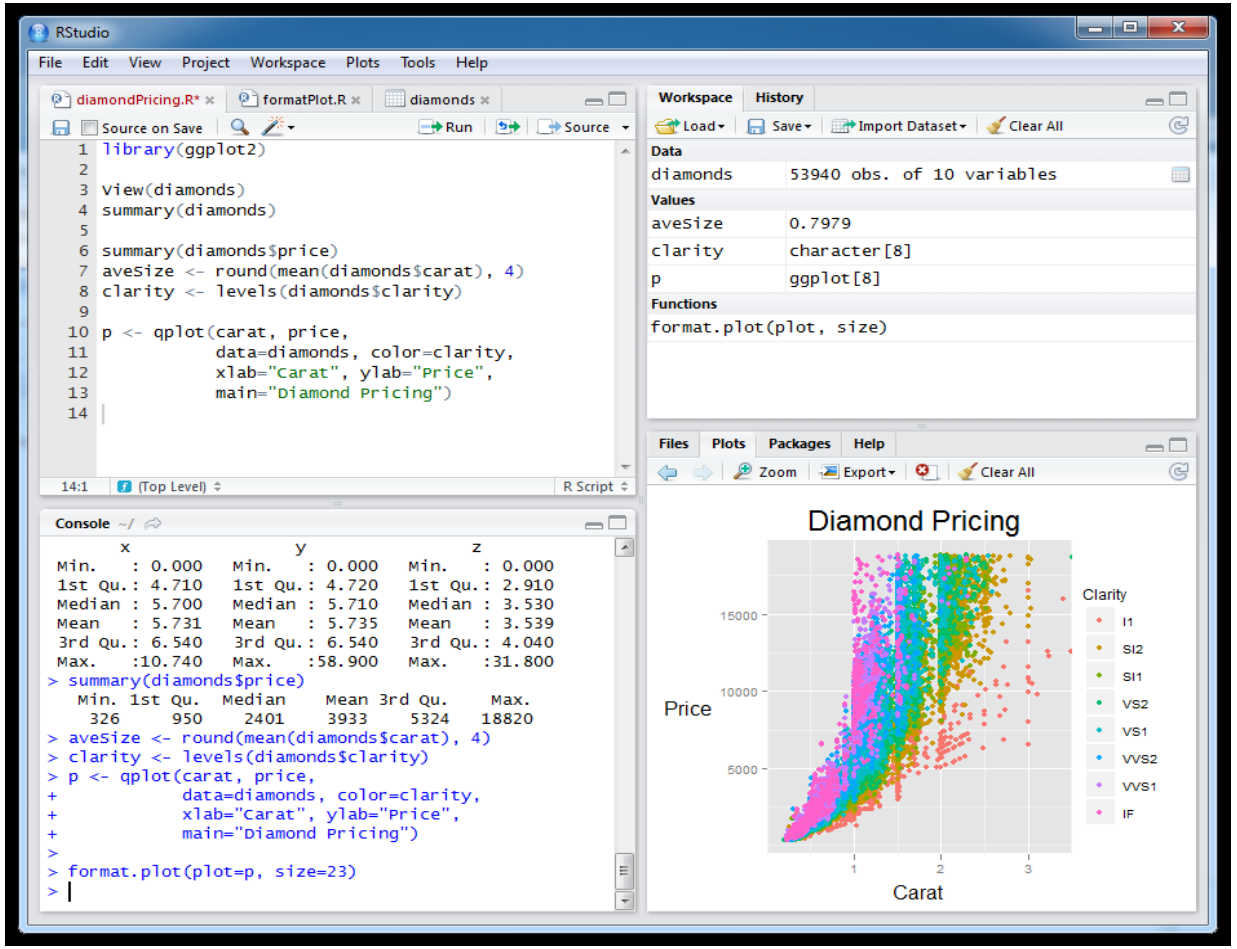
- Một “add-on” của R
- Website RStudio http://rstudio.org/

Giới thiệu RStudio

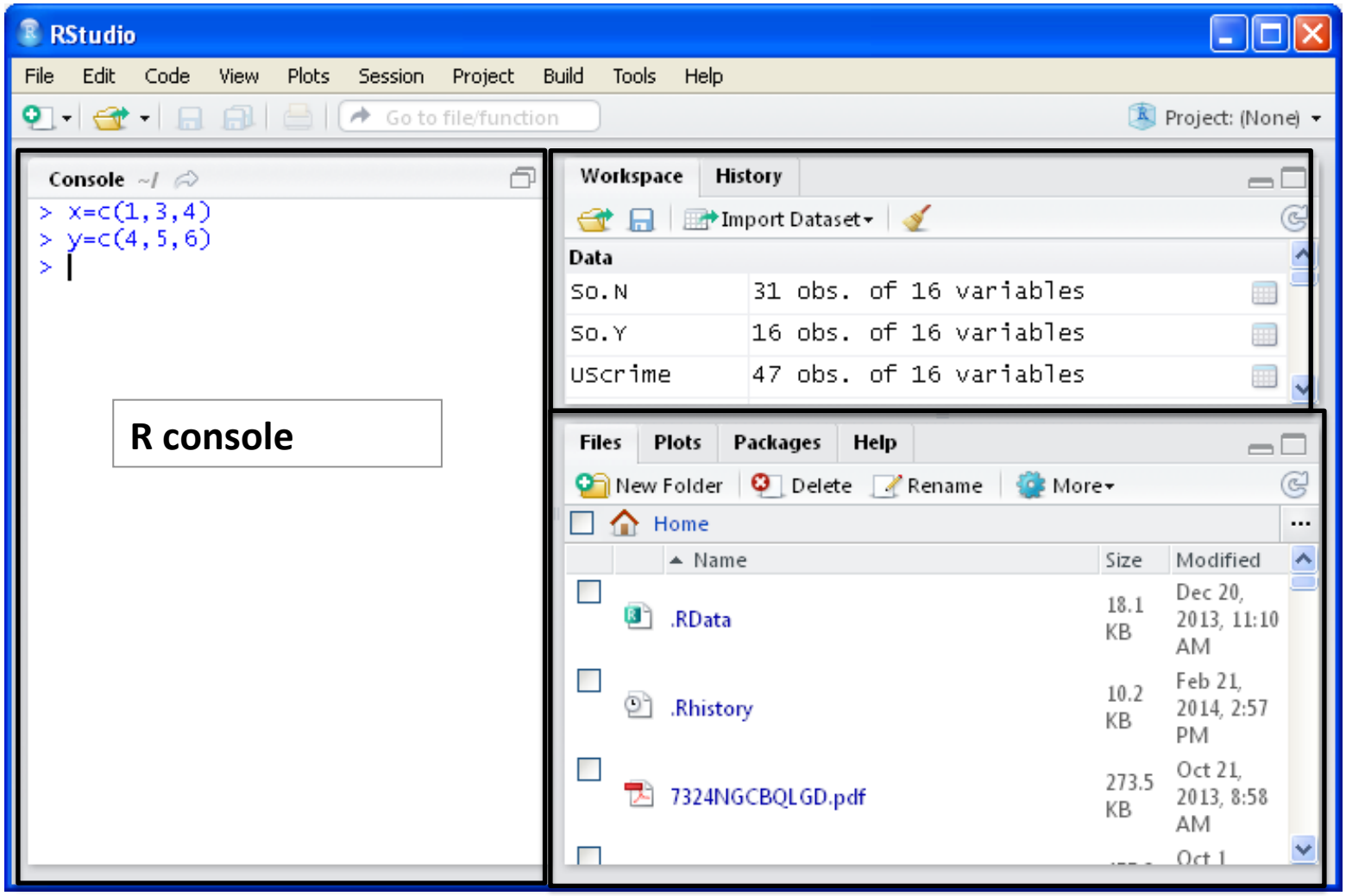
- Là một IDE (Interface Development Environment) của R.
- Cung cấp giao diện và một số tính năng để dễ dùng R hơn.
- Ngoài ra còn nhiều IDE khác:
    - TinnR
    - R commander

R và RStudio

- Cài đặt R trước
- Cài đặt RStudio



Màn hình RStudio



Workspace: Variables

File trên máy tính

Tóm lược

- R là một trong những phát triển quan trọng của khoa học thống kê
- Hoàn toàn miễn phí
- Sử dụng rộng rãi trong các đại học trên thế giới
- R vận hành theo **packages**
- **RStudio** là một “add-on” nhưng vận hành gần như độc lập với R