

Bài giảng 3: Biên tập dữ liệu

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia
Đại học Tôn Đức Thắng, Việt Nam

Nội dung

- Tạo biến mới
- Operators (hàm tính toán)
- Hàm có sẵn
- Sắp xếp dữ liệu – Sorting
- Hoán chuyển dữ liệu

Tạo biến mới

Arithmetic operators

Hàm	Mô tả
+	addition – cộng
-	subtraction – trừ
*	multiplication – nhân
/	division – chia
^ or **	exponentiation – lũy thừa
x %% y	modulus (x mod y) 5%%2 is 1
x %/% y	integer division 5%/2 is 2

Logical operators

Hàm

Mô tả

<

less than

<=

less than or equal to

>

greater than

>=

greater than or equal to

==

exactly equal to

!=

not equal to

!x

Not x

x | y

x OR y

x & y

x AND y

isTRUE(x)

test if x is TRUE

Hàm (số)

Hàm

abs(x)

sqrt(x)

**cos(x), sin(x),
tan(x)**

log(x)

log10(x)

exp(x)

Mô tả

absolute value

square root

also acos(x), cosh(x),
acosh(x), etc.

natural logarithm

common logarithm

e^x

Tạo biến mới

- Dataframe có tên: `tuan`
- Nếu dataframe có 2 cột (biến) `x1` và `x2`

```
x1 = c(1, 3, 4, 7)
```

```
x2 = c(4, 6, 8, 3)
```

```
tuan = data.frame(x1, x2)
```

Tạo biến mới

```
x1 = c(1, 3, 4, 7)
```

```
x2 = c(4, 6, 8, 3)
```

```
tuan = data.frame(x1, x2)
```

- Chúng ta có thể tạo ra một biến mới là tổng số của 2 cột:

```
sum = x1 + x2
```

- Nhưng biến sum sẽ không có trong **tuan** !

```
tuan
```


Giới thiệu dấu \$

- Dấu \$ dùng để tạo biến mới và kết nối với một dataframe.

```
tuan$sum = tuan$x1 + tuan$x2
```

```
tuan
```

Bây giờ dataframe tuan có 3 biến: x1, x2 và sum

Tạo biến mới qua mã hoá

Tạo biến mới qua coding

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

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

Mục tiêu: tạo ra biến mới là sex với 1=male, 2=female

Tạo biến mới qua coding

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

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

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

Mục tiêu: tạo ra biến mới là sex với 1=male, 2=female

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

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

Tạo biến mới qua coding

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

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

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

Mục tiêu: tạo ra biến mới là group

nếu id = 1,2,3 thì group = "A"

nếu id = 4,5 thì group = "B"

Tạo biến mới qua coding

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

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

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

Biến mới là group

```
dat$group[id >=1 & id <= 3] <- "A"
```

```
dat$group[id>=4 & id <=5] <- "B"
```

Hoán chuyển

Hoán chuyển dữ liệu

- Chuyển từ numeric sang text / character
- `as.numeric()`, `as.character()`, `as.vector()`, `as.matrix()`,
`as.data.frame`

ví dụ:

```
id1 = as.character(id)
```


Dataframe như là một matrix

Data frame = matrix

```
id = c(1:10)
```

```
name =
```

```
  c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J")
```

```
x = c(12, 45, 67, 32, 26, 86, 11, 16, 25, 37)
```

```
dat = data.frame(id, name, x)
```

Data frame là một ma trận (matrix)

Ma trận có dòng (row) và cột (column)

Data frame = matrix

```
dat = data.frame(id, name, x)
```

dat có 10 dòng và 3 cột

```
dat[,1]
```

```
dat[,2:3]
```

```
dat[2,]
```

```
dat[3:8,]
```

```
dat[1:5, 2:3]
```

```
> dat
```

	id	name	x
1	1	A	12
2	2	B	45
3	3	C	67
4	4	D	32
5	5	E	26
6	6	F	86
7	7	G	11
8	8	H	16
9	9	I	25
10	10	J	37

subset
làm việc với một phần dữ liệu

subset

**Chúng ta muốn làm việc
trong nhóm ID ≤ 5**

```
dat2 = subset(dat, id<=5)
```

```
dat3 = subset(dat, id<=8 &  
  x<30)
```

```
> dat
```

	id	name	x
1	1	A	12
2	2	B	45
3	3	C	67
4	4	D	32
5	5	E	26
6	6	F	86
7	7	G	11
8	8	H	16
9	9	I	25
10	10	J	37

order

Thứ tự hóa dữ liệu

Sắp xếp thứ tự - order

- Lệnh căn bản

order (variable)

- Từ cao xuống thấp (descending)

order (-variable)

Sắp xếp thứ tự: ví dụ

```
id = c(1:10)
name =
  c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J")
x = c(12, 45, 67, 32, 26, 86, 11, 16, 25, 37)
dat = data.frame(id, name, x)
```

sắp xếp thứ tự theo biến x

```
new.dat = dat[order(x), ]
```

hoặc

```
new.dat = dat[order(dat$x), ]
```


Sắp xếp thứ tự: ví dụ

sắp xếp thứ tự theo biến x, cao xuống thấp

```
new.dat = dat[-order(dat$x), ]
```

merge
Hợp nhất dữ liệu

Ví dụ một data frame

```
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)
```

melt(reshape)
chuyển từ cột sang dòng

Ví dụ một data frame

```
id=c(1:4)
sex=c("M","F","M","F")
group=c(1,1,2,2)
day1=c(15,16,21,31)
day2=c(17,15,23,35)
day3=c(19,20,19,33)
dat=data.frame(id,sex,group,day1,day2,day3)
dat
```

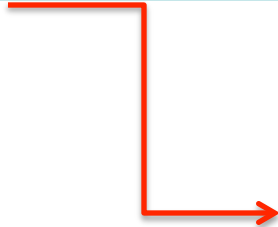
```
> dat
```

	id	sex	group	day1	day2	day3
1	1	M	1	15	17	19
2	2	F	1	16	15	20
3	3	M	2	21	23	19
4	4	F	2	31	35	33

Ví dụ một data frame

```
> dat
```

	id	sex	group	day1	day2	day3
1	1	M	1	15	17	19
2	2	F	1	16	15	20
3	3	M	2	21	23	19
4	4	F	2	31	35	33



	id	sex	group	day	value
1	1	M	1	day1	15
	1	M	1	day2	16
	1	M	1	day3	21
2	2	F	1	day1	16
	2	F	1	day2	15
	2	F	1	day3	20

melt trong gói reshape hay reshape2

```
id=c(1:4)
sex=c("M", "F", "M", "F")
group=c(1,1,2,2)
day1=c(15,16,21,31)
day2=c(17,15,23,35)
day3=c(19,20,19,33)
dat=data.frame(id,sex,group
               ,day1,day2,day3)
dat
```

```
require(reshape2)
dat1 = melt(dat,
            id=c("id", "sex", "group"),
            measure.vars=c("day1",
                           "day2", "day3"))
```

```
dat1 = melt(dat, id=1:3,
            measure.vars = c("day1",
                              "day2", "day3"))
```

melt trong gói reshape hay reshape2

```
require(reshape2)
dat1 = melt(dat,
  id=c("id", "sex", "group"),
  measure.vars=c("day1", "day2", "day3"))
```

```
> dat1
```

	id	sex	group	variable	value
1	1	M	1	day1	15
2	2	F	1	day1	16
3	3	M	2	day1	21
4	4	F	2	day1	31
5	1	M	1	day2	17
6	2	F	1	day2	15
7	3	M	2	day2	23
8	4	F	2	day2	35
9	1	M	1	day3	19
10	2	F	1	day3	20
11	3	M	2	day3	19
12	4	F	2	day3	33

cast(reshape)
chuyển từ dòng sang cột

Long data ...

```
> dat1
```

	id	sex	group	variable	value
1	1	M	1	day1	15
2	2	F	1	day1	16
3	3	M	2	day1	21
4	4	F	2	day1	31
5	1	M	1	day2	17
6	2	F	1	day2	15
7	3	M	2	day2	23
8	4	F	2	day2	35
9	1	M	1	day3	19
10	2	F	1	day3	20
11	3	M	2	day3	19
12	4	F	2	day3	33

Chuyển sang cột ...

```
> dat1
```

	id	sex	group	variable	value
1	1	M	1	day1	15
2	2	F	1	day1	16
3	3	M	2	day1	21
4	4	F	2	day1	31
5	1	M	1	day2	17
6	2	F	1	day2	15
7	3	M	2	day2	23
8	4	F	2	day2	35
9	1	M	1	day3	19
10	2	F	1	day3	20
11	3	M	2	day3	19
12	4	F	2	day3	33

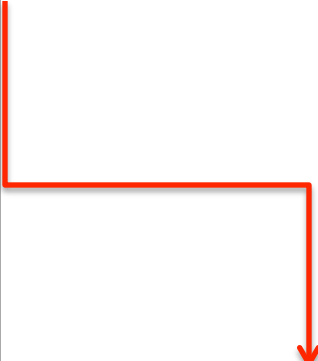
```
> dat
```

	id	sex	group	day1	day2	day3
1	1	M	1	15	17	19
2	2	F	1	16	15	20
3	3	M	2	21	23	19
4	4	F	2	31	35	33

Chuyển sang cột ...

```
> dat1
```

	id	sex	group	variable	value
1	1	M	1	day1	15
2	2	F	1	day1	16
3	3	M	2	day1	21
4	4	F	2	day1	31
5	1	M	1	day2	17
6	2	F	1	day2	15
7	3	M	2	day2	23
8	4	F	2	day2	35
9	1	M	1	day3	19
10	2	F	1	day3	20
11	3	M	2	day3	19
12	4	F	2	day3	33



```
dat2 = cast(dat1,  
            id+sex+group ~ variable)
```

Tóm tắt

- Coding: `dat$sex[gender=="male"] <- 1`
- Sort dữ liệu: **order()**
- Hợp nhất 2 dataframe: **merge**
- Chuyển từ cột sang dòng: **melt** (trong `reshape` hay `reshape2`)
- Chuyển từ dòng sang cột: **cast** (trong `reshape` hay `reshape`)