통계데이터베이스 및 실습(Part 4)
R을 이용한 데이터 조작

Jinseog Kim
Dep. of Applied Statistics, Dongguk University
Email: jinseog.kim @ gmail.com

2015-2

1. 학습목표

학습목표

1. 통계소프트웨어 R을 이용한 다양한 데이터 조작을 할 수 있다.
2. MySQL 쿼리문을 이용하여 데이터베이스 자료를 핸들링할 수 있다.
3. 통계소프트웨어 R에서 Excel 및 MySQL 데이터베이스 자료를 분석할 수 있다.
4. 빅데이터의 개념을 이해하고 R과 데이터베이스를 이용하여 빅데이터를 분석할 수 있다.

2. R을 이용한 데이터 조작

2.1 데이터 프레임(DATA FRAME)

R 데이터프레임

데이터 프레임은 "data.frame" 으로 분류되는 특별한 리스트로서, 리스트에 아래의 제약을 주어 만든다.

- 데이터프레임의 구성요소는 반드시 벡터, 펙터(factor), 행렬, 리스트 또는 다른 데이터프레임 이어야한다. (이 경우 모두 데이터프레임이 가능하다는 말은 아님)
- 행렬, 리스트 그리고 데이터 프레임이 가진 각각의 행, 구성요소 또는 변수는 새로운 데이터프레임의 행, 구성요소 또는 변수가 된다.
- 벡터(숫자, 문자등)는 데이터프레임의 하나의 열이 된다.
- 데이터 프레임에 포함된 변수는 그 길이가 모두 동일해야 한다.

데이터프레임 만들기

데이터프레임을 만들때는 data.frame() 함수를 이용한다.

```
> name <- c("kim","lee","park","Oh")
> sex <- c('f','m','f','m')
> income <- c(100,102,300,204)
> d1 <- data.frame(name=name, gender=sex, incom=income)
> d1
```

```
  name gender incom
1  kim      f   100
2  lee      m   102
3 park      f   300
4   Oh      m   204
```

리스트나 행렬을 통째로 데이터프레임으로 바꾸기 위해서는 as.data.frame()함수를 사용할 수 있다.

데이터프레임 조작

- 앞줄보기

```
> head(d1)
```

```
  name gender incom
1  kim      f   100
2  lee      m   102
3 park      f   300
4   Oh      m   204
```

- 변수명 출력

```
> names(d1)
```

```
[1] "name"   "gender" "incom"
```

- 행, 열의 수(차원)출력

```
> nrow(d1)
```

```
[1] 4
```

```
> ncol(d1)
```

```
[1] 3
```

```
> dim(d1)
```

```
[1] 4 3
```

SUBSCRIPTING(데이터에서 일부분을 추출)

- USArrests data :
  This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
- A data frame with 50 observations on 4 variables.
  1. Murder: numeric Murder arrests (per 100,000)
  2. Assault: numeric Assault arrests (per 100,000)
  3. UrbanPop: numeric Percent urban population
  4. Rape: numeric Rape arrests (per 100,000)

```
> head(USArrests,3)
```

```
        Murder Assault UrbanPop Rape
Alabama   13.2     236       58 21.2
Alaska    10.0     263       48 44.5
Arizona    8.1     294       80 31.0
```

SUBSCRIPTING

- Numeric subscripts

```
> # Top 5 states with high murder rate
> nidx <- order(USArrests$Murder, decreasing=T)[1:5]
> nidx
[1] 10 24  9 18 40
> USArrests[nidx,]
               Murder Assault UrbanPop Rape
Georgia          17.4     211       60 25.8
Mississippi      16.1     259       44 17.1
Florida          15.4     335       80 31.9
Louisiana        15.4     249       66 22.2
South Carolina   14.4     279       48 22.5
```

SUBSCRIPTING

- Logical subscripts

```
> lidx <- (USArrests$Murder
+           < quantile(USArrests$Murder, 0.1))
> head(lidx, 10)
```

```
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [8] FALSE FALSE FALSE
```

```
> USArrests[lidx,]
```

	Murder	Assault	UrbanPop	Rape
Iowa	2.2	56	57	11.3
Maine	2.1	83	51	7.8
New Hampshire	2.1	57	56	9.5
North Dakota	0.8	45	44	7.3
Vermont	2.2	48	32	11.2

SUBSET

- subset함수

```
> subset(USArrests, UrbanPop > 85)
```

```
             Murder Assault UrbanPop Rape
California      9.0     276       91 40.6
New Jersey      7.4     159       89 18.8
New York       11.1     254       86 26.1
Rhode Island    3.4     174       87  8.3
```

```
> subset(USArrests, UrbanPop < 40 &  Murder < 10,
+        select = c(Assault, Rape))
```

```
              Assault Rape
Vermont            48 11.2
West Virginia      81  9.3
```

2.2 DATA.TABLE을 이용한 데이터 핸들링

DATA.TABLE 객체생성

data.table(...)

```
> library(data.table)
> DT <- data.table(x=c("b","b","a","a"),v=rnorm(4))
> DT
```

```
   x           v
1: b -1.545448388
2: b -0.528393243
3: a -1.086758791
4: a -0.000111512
```

DATA.FRAME으로부터 DATA.TABLE 객체생성

```
> CARS <- data.table(cars)
> head(CARS)
```

```
   speed dist
1:     4    2
2:     4   10
3:     7    4
4:     7   22
5:     8   16
6:     9   10
```

DATA.TABLE 목록

```
> tables()
     NAME NROW NCOL MB COLS        KEY
[1,] CARS   50    2  1 speed,dist
[2,] DT      4    2  1 x,v
Total: 2MB
```

Group summary

```
> Iris <- data.table(iris)
> names(Iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length"
[4] "Petal.Width"  "Species"
> Iris[, mean(Petal.Width), by="Species"]
      Species    V1
1:     setosa 0.246
2: versicolor 1.326
3:  virginica 2.026
> Iris[,lapply(.SD, mean),by=Species]
      Species Sepal.Length Sepal.Width
1:     setosa        5.006       3.428
2: versicolor        5.936       2.770
3:  virginica        6.588       2.974
   Petal.Length Petal.Width
1:        1.462       0.246
2:        4.260       1.326
3:        5.552       2.026
> tapply(iris$Petal.Width, iris$Species, mean)
    setosa versicolor  virginica
     0.246      1.326      2.026
```

R 데이터 객체

또한 위 자료에서 고등학교 학생들의 1999년 API성적과 2000년 API성적에 대하여 대응비교(t test)를 수행한 것이다.

```
t.test(b$api100, b$api99, paired=T,
       alternative="greater")
```

```
        Paired t-test

data:  b$api100 and b$api99
t = 15.2003, df = 754, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 11.36122    Inf
sample estimates:
mean of the differences
               12.74172
```