데이터 프레임

Jeon Jong-June

Monday, March 09, 2015

데이터 프레임 만들기

데이터 프레임은 데이터 테이블을 저장하는 R의 대표적인 데이터 형식이다. data.frame 함수를 이용하여 생성한다.

```
kids <- c("Jack", "Jill")
ages <- c(12, 10)
d <- data.frame(kids, ages, stringsAsFactors = F)
d
```

```
##   kids ages
## 1 Jack   12
## 2 Jill   10
```

```
str(d)
```

```
## 'data.frame':    2 obs. of  2 variables:
##  $ kids: chr  "Jack" "Jill"
##  $ ages: num  12 10
```

stringsAsFactors 는 data.frame 함수의 option 으로 문자형 변수를 R에서 정의한 팩터라는 변수형식으로 변환 여부를 결정한다. F라 하면 문자형 변수를 팩터 형식으로 변환하지 않는다.

접근하기

```
d$ages
```

```
## [1] 12 10
```

```
class(d$ages)
```

```
## [1] "numeric"
```

```
names(d)
```

```
## [1] "kids" "ages"
```

데이터 프레임은 행렬과 같은 방식으로 행과 열의 index를 통해 접근할 수 있다

```r
d[,1:2]
```

```
##   kids ages
## 1 Jack   12
## 2 Jill   10
```

```r
class(d[,1:2])
```

```
## [1] "data.frame"
```

파일 다루기

파일읽기

read.table 함수를 이용한다.

```r
read.table(file, header = FALSE, sep = "", stringsAsFactors= F)
```

```r
A <-read.table("C:/Users/uos_stat/Documents/CO2.dat", header = TRUE, sep = " ", stringsAsFactors= F)
head(A)
```

```
##   Plant   Type  Treatment conc uptake
## 1   Qn1 Quebec nonchilled   95   16.0
## 2   Qn1 Quebec nonchilled  175   30.4
## 3   Qn1 Quebec nonchilled  250   34.8
## 4   Qn1 Quebec nonchilled  350   37.2
## 5   Qn1 Quebec nonchilled  500   35.3
## 6   Qn1 Quebec nonchilled  675   39.2
```

```r
class(A$Plant)
```

```
## [1] "character"
```

```r
A <-read.table("C:/Users/uos_stat/Documents/CO2.dat", header = TRUE, sep = " ", stringsAsFactors= T)
head(A)
```

```
##   Plant   Type  Treatment conc uptake
## 1   Qn1 Quebec nonchilled   95   16.0
## 2   Qn1 Quebec nonchilled  175   30.4
## 3   Qn1 Quebec nonchilled  250   34.8
## 4   Qn1 Quebec nonchilled  350   37.2
## 5   Qn1 Quebec nonchilled  500   35.3
## 6   Qn1 Quebec nonchilled  675   39.2
```

```r
class(A$Plant)
```

```
## [1] "factor"
```

문자열로 저장된 Plant 변수의 경우 “stringsAsFactors= T” 의 옵션일때 팩터 형태로 나타난다.

파일 쓰기

```
head(USArrests)
```

```
##            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
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
class(A)
```

```
## [1] "data.frame"
```

```
A <- write.table(USArrests , file = "C:/Users/uos_stat/Documents/US.csv",
            sep = ",",row.names = FALSE,col.names = TRUE)
```

sep에 옵션에 따라 구분자를 변경할 수 있다. row.names가 TRUE인 경우에 데이터의 rownames이 함께 저장된다. col.name 역시 마찬가지다.

데이터프레임의 결합

rbind와 cbind 함수 사용하기

데이터 프레임 역시 행렬과 마찬가지로 rbind와 cbind를 사용할 수 있다.

```
A = data.frame(x1 = rep(0,10), x2 = rep('b',10))
B = data.frame(x3 = rep(1,10), x2 = rep('d',10))
AB = cbind(A,B)
head(AB)
```

```
##   x1 x2 x3 x2
## 1  0  b  1  d
## 2  0  b  1  d
## 3  0  b  1  d
## 4  0  b  1  d
## 5  0  b  1  d
## 6  0  b  1  d
```

rbind(A,B) 작동하지 않는 것을 확인해라.

Merge

두개의 데이터 프레임을 결합하고자 할 때 사용하는 명령어다. 다음예제를 살펴보자

```
d1 = data.frame(kids = c("Jack", "Jill", "Jillian", "John"),
                states = c("CA", "MA", "MA", "HI"))

d2 <- data.frame(ages = c(10, 7, 12), kids = c("Jill", "Jillian", "Jack") )
d1
```

```
##      kids states
## 1    Jack     CA
## 2    Jill     MA
## 3 Jillian     MA
## 4    John     HI
```

```
d2
```

```
##   ages    kids
## 1   10    Jill
## 2    7 Jillian
## 3   12    Jack
```

```
d <-merge(d1, d2)
d
```

```
##      kids states ages
## 1    Jack     CA   12
## 2    Jill     MA   10
## 3 Jillian     MA    7
```

위 예제는 kids의 변수정보를 이용하여 대아토거 합쳐진것이다. 두 데이터 프레임이 공통적으로 가진 변수를 이용해서 데이터를 합치는 명령어가 merge다.

```
d3 <- data.frame(ages = c(10, 7, 12), pals = c("Jill", "Jillian", "Jack") )
d <-merge(d1, d3, by.x = 'kids', by.y = "pals")
d
```

```
##      kids states ages
## 1    Jack     CA   12
## 2    Jill     MA   10
## 3 Jillian     MA    7
```

위와 같이 by.x 와 by.y 를 이용해서 데이터를 합칠때 사용할 기준변수를 정해줄수 있다.

```
d <-merge(d1, d3, by.x = 'kids', by.y = "pals", all.x = TRUE)
d
```

```
##      kids states ages
## 1    Jack     CA   12
## 2    Jill     MA   10
## 3 Jillian     MA    7
## 4    John     HI   NA
```

all.x 혹은 all.y를 이용해서 어떤 한쪽변수가 완전히 출력되도록 할 수 있다. all = T 라고 하면 by.x 혹은 by.y 를 통해 대응되지 않는 변수가 모두 출력된다.