Seok-Oh Jeong, In Heok Lee, and Jay W. Rojewski
University of Georgia

The *R* Workshop

Applying the Integrated Suite of Software Facilities for Statistical Computing and Graphics

University of Georgia
Department of Workforce Education, Leadership, and Social Foundations
College of Education Research Office

January 23-January 24, 2012

January 23-January 24, 2012

2. Data Manipulation

University of Georgia
Department of Workforce Education, Leadership, and Social Foundations
College of Education Research Office

Data Types

- Vector

```
x <- c(1,2,3,4,5)
> y <- c("Clarke", "Oconee", "Barrow")
> x[3]
```

- Array

```
> z <- array(1:20, dim=c(4,5))
> z[4,5]
```

- Matrix

```
> z <- matrix(1:20, 4, 5)
> A <- matrix(2, 4, 5)
> z[3,2]
```

Data Manipulation

Data Types

- List

```
> Jeong <- list(first.name="Seok-Oh",
age=40, citizenship="South Korea")
```

- Data frame

```
> X <- c("Jeong", "Lee", "Rojewski")
> Y <- c(40, 35, 53)
> data <- data.frame(last.names=X,
age=Y)
```

- Factor

```
> z <- c("LD","BD", "BD", "LD", "Non",
"Non", "Non", "Non")
> z <- factor(z)
> levels(z)
```

Data Manipulation

Vectors

- Concatenation

```
> a <- c(2,2,2,2,2,2)
> a <- c(1, 2, 3); b <- c(5, 6)
> x <- c(a, 4, b) # x <- c(1,2,3,4,5,6)
```

- Sequence

```
> x <- seq(from=0, to=1, by=0.1)
> y <- seq(from=0, to=1, length=11)
> z <- 1:10
> rep(1, 10)
```

Vectors

- Arithmetic: Componentwise

```
> x <- 1:3; y <- c(2,2,2)
> x+y
> x-y
> x*y
> x/y
> x^y
> z <- rep(2, 5)
> x+z
> y-3
```

Vectors

- Mathematical functions

```
> x <- 1:10          > log(x)
> exp(x)             > sin(x)   # cos(x), tan(x)
> abs(x)             > sqrt(x)
> sort(x)            > length(x)
> min(x)             > max(x)
> mean(x)            > sum(x)
> prod(x)
```

Vectors

- Logical vectors

```
> x <- 1:10; y <- rep(5, 10)
> z <- x<5              # less than
> sum(z)
> x<=5                  # less than or equal to
> x==5                  # equal
> x!=5                  # not equal
> (x>5)&(y<2)           # and
> (x<5)|(y<2)           # or
```

- Missing values

```
> x <- c(1, 2, 3, NA, 5)
> is.na(x)
```

Vectors

- Index vectors

```
> x <- -10:10
> x[3]
> x[1:3]
> x[c(1,3,5)]
> y <- x[x<0]
> x[x<0] <- -x[x<0]
> x <- c(1, 2, 3, NA, 5)
> x[!is.na(x)]
> x[is.na(x)] <- 4
> fruit <- c(5, 3, 2)
> names(fruit) <- c("apple", "orange", "peach")
> fruit[c("apple", "peach")]
```

Arrays and Matrices

- To generate an array and a matrix

```
> z <- array(1:20, dim=c(4,5))
> A <- matrix(1:20, 4, 5)
> B <- matrix(2, 4, 5)
> z[3,4]
> A[3,4]
> x <- c(1,2,3)
> y <- c(4,5,6)
> cbind(x, y)
> rbind(x, y)
> cbind(B, 1:4)
> C <- cbind(A, B)
```

Arrays and Matrices

- Arithmetic: Componentwise

```
> A <- matrix(1:20, 4, 5)
> B <- matrix(1:20, 4, 5)
> A+B
> A-B
> A*B
> A/B
```

- Arithmetic: Matrix multiplication, inverse

```
> A <- matrix(runif(20), 4, 5)
> B <- A%*%t(A)      # t(): transpose
> solve(B)           # inverse
```

Lists

- A list is an object consisting of a collection of objects called components.

```
> Jeong <- list(first.name="Seok-Oh", age=40,
  married=T, no.children=2, child.ages=c(9, 6))
> Jeong$age
> Jeong[[1]]
> Jeong$child.ages
> Jeong[[5]][1]
```

Factors

- A factor is a vector object used to specify a discrete classification (grouping) of the components of other vectors of the same length.

```
> x <- c(80, 90, 85, 85, 50, 60, 45, 50)
> z <- c("LD","BD", "BD", "LD", "Non", "Non",
"Non", "Non")
> z <- factor(z)
> levels(z)
> x.means <- tapply(x, z, mean)
```

Data frames

- A data frame is a list with restrictions:
  a. The components must be vectors (numeric, character, or logical), factors, numeric matrices, lists, or other data frames.
  b. Numeric vectors, logicals and factors are included as is, and character vectors are coerced to be factors, whose levels are the unique values appearing in the vector.
  c. Vector structures appearing as variables of the data frame must all have the same length, and matrix structures must all have the same row size.