Data Management I

Nathaniel MacNell

EPID 799B

Overview

- 3 Lectures on data management
  - Today: single-variable
  - Next: whole-dataset
  - Finally: advanced stuff

- Today:
  - Factors
  - Activity

Factors: category variables in R

- Similar to SAS data labels
- Factor **levels:** ordered values of the factor (1,2,3,4)
- Factor **labels:** the text value for each value (“low”)

```
x <- c(1,2,1,1,1,2)   # make some data
xf <- factor(x,       # make factor
        levels=c(1,2),
        labels=c("f","male") )
```

What's the point of factors?

- Many functions want or use properties of factors:
  - glm (regression) needs categorical variables as factors
  - Many plotting functions use factor levels for ordering
- Ease & flexibility comes at a price
  - Some functions interpret the factor **levels** or alternately, the **labels**:

Level	1	2	3
Label	40	50	60

```
as.numeric(x)                          [1] 1 2 3
as.numeric( as.character(x) )          [1] 40 50 60
```

Basic Data Management Tasks: []

- Remember that [] can be used to “subset” the destination (right side), not just the source (right side). This is how you selectively modify variables!

```
age[age>65] <- 65
age[is.na(age)] <- mean(age,na.rm=TRUE)

names(births)[2] <- “newname”
```

Activity

- Small groups (posted on GitHub)
- 4 Sections
  - Factors
  - Dates
  - Loops
  - Data transformation