Data Manipulation

Data Recoding

Recoding

- One data manipulation task that you need to do in pretty much any data analysis is recode data

- Replacing data in an existing field or recoding into a new field based on criteria you specify

- Recoding is also known as **Replacing** or **Imputation**

Recoding the existing values to new values

• **Ex:**

x <- c(3,4,5,6,7,8)   # numeric vector

Recode the values less than 6 with zero

> x[x<6] <- 0

> x

[1] 0 0 0 6 7 8

Write a command to recode the values between 4 and 8 with 100 in the vector x

Recoding Values to Missing

- Values of a vector or data frame can be recoded to **NA** if required

Ex:

x <- c(3,4,5,6,7,8)

> x[x==6] <- NA

> x

[1] 3 4 5 NA 7 8

Write a command to recode the values greater than 6 with NA in the vector x

Recoding missing values by another value

> A <- c(3, 2, NA, 5, 3, 7, NA, NA, 5, 2, 6)

> A

o/p: [1] 3 2 NA 5 3 7 NA NA 5 2 6

We can re-code all missing values by another number (such as zero) as follows:

> A[is.na(A)] <- 0

> A

o/p: [1] 3 2 0 5 3 7 0 0 5 2 6

Recoding a categorical variable or factor

- re-coding tasks are more complex, particularly when you wish to re-code a categorical variable or factor

Ex:

gender <- c("MALE","FEMALE","FEMALE","MALE","MALE")

> gender

o/p: [1] "MALE" "FEMALE" "FEMALE" "MALE " "MALE"

> gender[gender=="MALE"]<- 1

> gender

o/p: [1] "1"   "FEMALE" "FEMALE" "1" "1"

Recoding a categorical variable or factor

Ex-2:

gender <- c("MALE","FEMALE","FEMALE","MALE","MALE")

> gender

o/p: [1] "MALE" "FEMALE" "FEMALE" "MALE " "MALE"

Recode MALE by 1 and FEMALE by 2 using ifelse()

> ifelse(gender == "MALE", 1, 2)

[1] 1 2 2 1 1

Recoding a categorical variable or factor

Ex-3:

gender <- c("MALE","FEMALE","FEMALE","UNKNOWN","MALE")

> gender

o/p: [1] "MALE"   "FEMALE" "FEMALE" "UNKNOWN" "MALE"

Recode MALE by 1, FEMALE by 2 and UNKNOWN by 3 using ifelse()

> gender <- c("MALE","FEMALE","FEMALE","UNKNOWN","MALE")

> ifelse(gender == "MALE", 1, ifelse(gender == "FEMALE", 2, 3))

[1] 1 2 2 3 1

Recoding values in Data Frame

> A <- data.frame(Gender = c("F", "F", "M", "F", "B", "M", "M"), Height = c(154, 167, 178, 145, 169, 183, 176))

> A

o/p:

	Gender	Height
1	F	154
2	F	167
3	M	178
4	F	145
5	B	169
6	M	183
7	M	176

Recoding values in Data Frame

This one gets re-coded to the value 99.

Note that the Gender variable is located in the first column, or A[,1]

A[,1] <- ifelse(A[,1] == "M", 1, ifelse(A[,1] == "F", 2, 99))

> A

o/p:

	Gender	Height
1	2	154
2	2	167
3	1	178
4	2	145
5	99	169
6	1	183
7	1	176

Recode Data in an existing field

 # Replace all the data in a field with a number

SchoolData$Grade <- 5

 # Replace all the data in a field with with text

SchoolData$Grade <- "Five"

 # Replace all the data in a field with NA (missing data)

SchoolData$Grade <- NA

Recode Data in an existing field

 # Replace the data in a field based on equal to some value

SchoolData$Grade[SchoolData$Grade==5] <- "Grade Five"

 # Or replace based on greater than or equal to some value

SchoolData$Grade[SchoolData$Grade<=5] <- "Grade Five or Less"

 # Or replace based on equal to some text

SchoolData$Grade[SchoolData$Grade=="Five"] <- "Grade Five"

Recode Data in an existing field

 # Or replace only missing data

 # Note that ==NA does not work!

SchoolData$Grade[is.na(SchoolData$Grade)] <- "Missing Grade"

Recode into a new field

First create the new column

SchoolData$CopyOfGrade <- NA

Then copy the data from the existing column into the new one.

SchoolData$CopyOfGrade <- SchoolData$Grade

Recode into a new field

Recode into a new field

Recode into a new field in R

First create the new field

StudentData$NewGrade <- NA

Then recode the old field into the new one for the specified rows

SchoolData$NewGrade[SchoolData$Grade==5] <- 5