Data analysis with R software

Data-analyysi R-ohjelmistolla

Tommi Härkänen

National Institute for Health and Welfare (THL), Helsinki
E-mail: tommi.harkanen@helsinki.fi

University of Helsinki, January 16, 2013

Summary of 1st lecture

- Basic syntax and user interface
- Data structures:
  - Scalars and vectors
  - Character strings
  - Data frames and matrices
  - Lists
- Function calls
- Graphical output
- Logical operations
- Assignments

Contents of 2nd lecture

- Missing values, 'NA'
- Transformations of data frames
- Ordering and sorting data
- Factor variables
- The NHANES data

National Health and Nutrition Examination Survey (NHANES)

- The NHANES has been designed to collect information about the health and diet of people in the United States.
- The Third National Health and Nutrition Examination Survey (NHANES III), 1988-94, contains data for 33,994 persons ages 2 months and older who participated in the survey.
- In this example 49 variables and 20,050 observations are being used.

Variables of NHANES

Race-ethnicity	Other than pregnant,doctor told diabetes
Race	Age first told you had diabetes – yrs
Ethnicity	Are you now taking insulin
Sex	How long since doctor took blood press
Age at interview (screener) - qty	Doctor ever told had hypertension/HBP
Family size (persons in family)	Told 2+ times you had hypertension/HBP
Household size (persons in dwelling)	Now taking prescribed medicine for HBP
FIPS code for State	Ever had blood cholesterol checked
Anyone living here smoke cigs in home	Doctor told blood cholesterol level high
Do you have enough food to eat	Take prescribed med to lower cholesterol
Doctor ever told you had: arthritis	Ever had any pain or discomfort in chest
Type arthritis:rheumatoid,osteoarthritis	Get chest pain when walk uphill or hurry
Doctor told: congestive heart failure	Get chest pain if walk at ordinary pace
Doctor ever told you had: stroke	Doctor ever told you had a heart attack
Doctor ever told you had: asthma	How many heart attacks have you had
Doctor ever told had: chronic bronchitis	Age when you had 1st heart attack – yrs
Age when first told you had arth – yrs	Age when had last heart attack – yrs
Age 1st told had cong heart fail – yrs	How tall are you without shoes - inches
Age when 1st told you had stroke – yrs	How much do you weigh w/out clothes -lbs
Did mother have diabetes	Have you smoked 100+ cigarettes in life
Did father have diabetes	Age when you started smoking regularly
Did mother have heart attack	Do you smoke cigarettes now
Did father have heart attack	# cigarettes smoked per day
Ever been told you have sugar/diabetes	How many yrs have you smoked this amount
Were you pregnant when told had diabetes	Ever period of 1+ years when smoked more

Missing values

```
x <- c(1, 2, NA, 4)
mean(x)
mean(x, na.rm=TRUE)
x[x == NA]
is.na(x)
x[!is.na(x)]
```

```
[1] NA

[1] 2.333

[1] NA NA NA NA

[1] FALSE FALSE  TRUE FALSE

[1] 1 2 4
```


Select subset of data frame

```
d <- data.frame(a=11:14, b=seq(0, 1, length=4), c=letters[1:4])
rownames(d) <- paste("Observation", 1:4)
d
subset(d, a < 14 & b >= 0.1, select=b:c)
```

```
              a      b c
Observation 1 11 0.0000 a
Observation 2 12 0.3333 b
Observation 3 13 0.6667 c
Observation 4 14 1.0000 d

                  b c
Observation 2 0.3333 b
Observation 3 0.6667 c
```


Transform variables of a data frame

```
d <- data.frame(a=11:14, b=seq(0, 1, length=4), c=letters[1:4])
rownames(d) <- paste("Observation", 1:4)
d
within(d, {idx <- b < 0.5; d <- NA; d[idx] <- a[idx] + b[idx];
           rm(idx)})
```

```
              a      b c
Observation 1 11 0.0000 a
Observation 2 12 0.3333 b
Observation 3 13 0.6667 c
Observation 4 14 1.0000 d

              a      b c     d
Observation 1 11 0.0000 a 11.00
Observation 2 12 0.3333 b 12.33
Observation 3 13 0.6667 c    NA
Observation 4 14 1.0000 d    NA
```

Merge data frames

```
d <- data.frame(a=11:14, b=seq(0, 1, length=4), c=letters[1:4])
rownames(d) <- paste("Observation", 1:4)
d3 <- data.frame(a=c(2:4), e=LETTERS[c(2:4)])
d3
merge(d, d3, all=TRUE)
```

```
  a e
1 2 B
2 3 C
3 4 D

   a      b    c    e
1  2     NA <NA>    B
2  3     NA <NA>    C
3  4     NA <NA>    D
4 11 0.0000    a <NA>
5 12 0.3333    b <NA>
6 13 0.6667    c <NA>
7 14 1.0000    d <NA>
```

Data analysis with R software

Factor variables

```
educ.yrs <- factor(c(9, 12, 17, 9, 17, 17), levels = c(9, 12,
    17), labels = c("basic", "middle", "high"))
educ.yrs

## [1] basic  middle high   basic  high   high
## Levels: basic middle high

## Combine two levels:
table(educ.yrs)

## educ.yrs
##  basic middle   high
##      2      1      3

educ.yrs[educ.yrs == "basic"] <- "middle"
table(educ.yrs)

## educ.yrs
##  basic middle   high
##      0      3      3
```