

Biostatistics

Math 322 — Spring 2009

Comments about Homework 1

Problem 3 (Problems 2.31 – 2.32 in the book)

The data set given in the file `LEAD.DAT` (in the `Data sets/ASCII - .dat tab` directory of the CD-ROM) is formatted so that it uses two lines per record. This confuses the `read.table`-function. An alternative is to use the more flexible `scan`-function, which merely reads all the numbers into a vector. We can then change the vector into a matrix or dataframe afterwards.

For instance,

```
p3.rawdata <- scan("LEAD.DAT")
p3.datamatrix <- matrix(p3.rawdata, ncol = 41, byrow = T)
```

Our data are now stored in the matrix `p3.datamatrix`. Individual columns (fields) can be accessed using the square bracket notation, for instance `p3.datamatrix[,4]` is a vector representing the 4th column (the age).

Another alternative is to use the comma separated data file `lead.txt` (in the `Data sets/ASCII - .txt comma` directory of the CD-ROM), where each record is kept on one line. To read in a comma separated data file, we use the `read.csv`-function, which works just as `read.table`.

For instance,

```
p3.dataframe <- read.csv("lead.txt")
```

This data file contains headers, so it has the added benefit that the columns automatically have meaningful names. To access individual columns of a data frame, use the dollar notation. For example `p3.dataframe$Age` is a vector representing the age.