Data Analysis with R - Day 3

ylee@dongguk.edu // dryl@icloud.com

2019-4-30

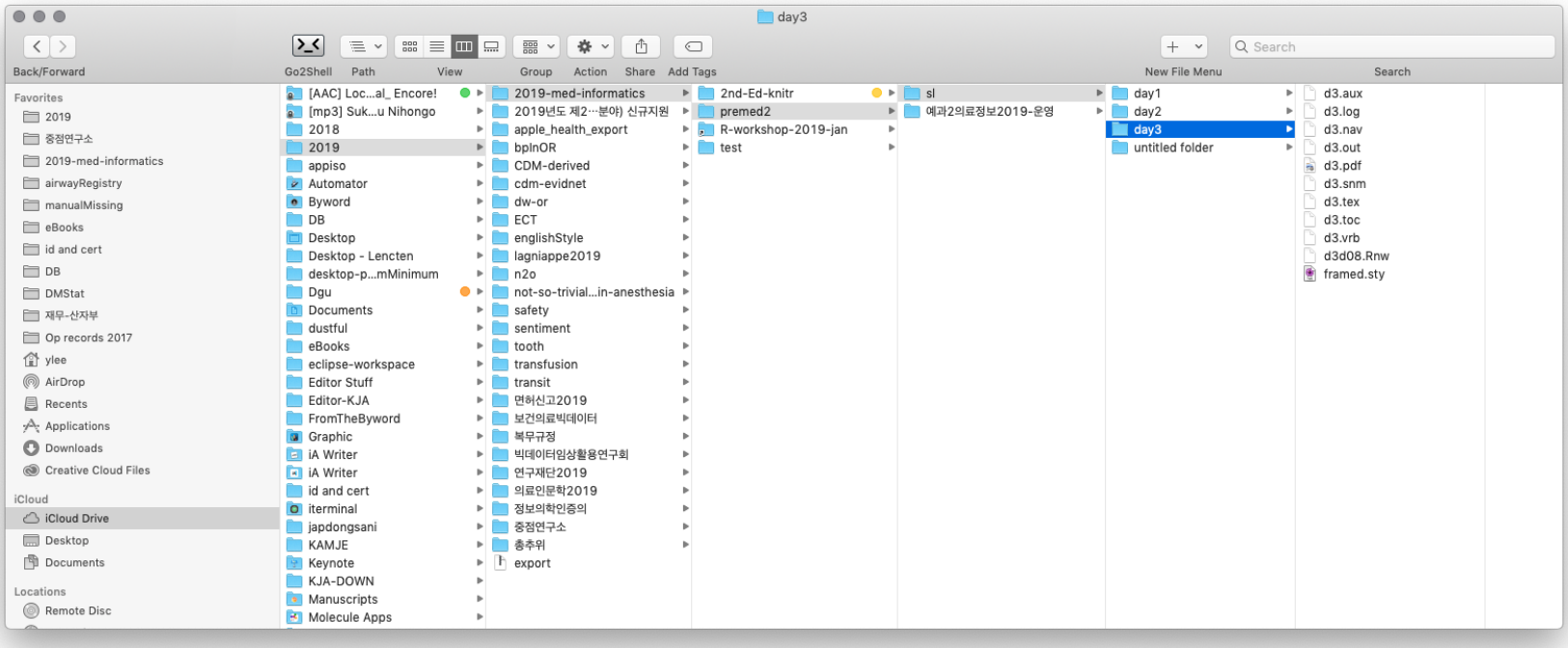
1. File I/O

2. Graphing

File I/O

File Input/Output

First, one must be able to locate a file.



In A Typical Unix Family Machine

Machine nickname: `eblis`

User name: `ylee`

Name	Display	Full Path
Home	`eblis: ylee$`	`/Users/ylee`
Root	`eblis:/ ylee$`	`/`
R Library	`eblis:library ylee$`	`/Library/Frameworks/R.framework/Versions/3.5/Resources/library`

Set up the working directory in R: **setwd()** and **getwd()**

```
(oldPath <- getwd())

## [1] "/Users/ylee/Library/Mobile Documents/com~apple~CloudDocs/2019/2019-med-informatics/premed2/sl/day3"

setwd("/Users/ylee")
getwd()

## [1] "/Users/ylee"

dir()

##  [1] "Applications"         "Creative Cloud Files" "Desktop"
##  [4] "Documents"            "Downloads"            "eclipse"
##  [7] "Library"              "Movies"               "Music"
## [10] "Pictures"             "Public"               "VirtualBox VMs"

setwd(oldPath)
getwd()

## [1] "/Users/ylee/Library/Mobile Documents/com~apple~CloudDocs/2019/2019-med-informatics/premed2/sl/day3"

dir()

##  [1] "boxplot.pdf"   "d3.aux"        "d3.log"        "d3.nav"
##  [5] "d3.out"        "d3.pdf"        "d3.snm"        "d3.tex"
##  [9] "d3.toc"        "d3.vrb"        "d3d08.Rnw"     "d3d09.Rnw"
## [13] "d3d095.Rnw"    "d3x.aux"       "d3x.log"       "d3x.nav"
## [17] "d3x.out"       "d3x.snm"       "d3x.tex"       "d3x.toc"
## [21] "d3x.vrb"       "d3xp.pdf"      "d3y.aux"       "d3y.log"
## [25] "d3y.nav"       "d3y.out"       "d3y.pdf"       "d3y.snm"
## [29] "d3y.tex"       "d3y.toc"       "d3y.vrb"       "figure"
## [33] "framed.sty"    "navig.nav"     "Orthe.RData"   "Orthodont.csv"
```

R data.frame: revisit

- A final data is grid-structured and stored *commonly* in **data.frame()** in R.
- R distributes enormous amount of educational material as its **data.frame()**.

	var1	var2	var3	var4	...
record1	...	...	...	...	...
record2	...	...	...	...	...
record3	...	...	...	...	...
.					
.					
.					

Try Google with "R common datasets to learn ..."

Orthodont dataset

```
data(Orthodont, package = "nlme")
is.data.frame(Orthodont)

## [1] TRUE

head(Orthodont, 5)

##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male

str(Orthodont)

## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 108 obs. of  4 variables:
##  $ distance: num  26 25 29 31 21.5 22.5 23 26.5 23 22.5 ...
##  $ age     : num  8 10 12 14 8 10 12 14 8 10 ...
##  $ Subject : Ord.factor w/ 27 levels "M16"<"M05"<"M02"<..: 15 15 15 15 3 3 3 3 7 7 ...
##  $ Sex     : Factor w/ 2 levels "Male","Female": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "outer")=Class 'formula'  language ~Sex
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##  - attr(*, "formula")=Class 'formula'  language distance ~ age | Subject
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##  - attr(*, "labels")=List of 2
##   ..$ x: chr "Age"
##   ..$ y: chr "Distance from pituitary to pterygomaxillary fissure"
##  - attr(*, "units")=List of 2
##   ..$ x: chr "(yr)"
##   ..$ y: chr "(mm)"
##  - attr(*, "FUN")=function (x)
##   ..- attr(*, "source")= chr "function (x) max(x, na.rm = TRUE)"
##  - attr(*, "order.groups")= logi TRUE
```

```
Orthodont
```

```
##     distance age Subject    Sex
## 1       26.0   8     M01   Male
## 2       25.0  10     M01   Male
## 3       29.0  12     M01   Male
## 4       31.0  14     M01   Male
## 5       21.5   8     M02   Male
## 6       22.5  10     M02   Male
## 7       23.0  12     M02   Male
## 8       26.5  14     M02   Male
## 9       23.0   8     M03   Male
## 10      22.5  10     M03   Male
## 11      24.0  12     M03   Male
## 12      27.5  14     M03   Male
## 13      25.5   8     M04   Male
## 14      27.5  10     M04   Male
## 15      26.5  12     M04   Male
## 16      27.0  14     M04   Male
## 17      20.0   8     M05   Male
## 18      23.5  10     M05   Male
## 19      22.5  12     M05   Male
## 20      26.0  14     M05   Male
## 21      24.5   8     M06   Male
## 22      25.5  10     M06   Male
## 23      27.0  12     M06   Male
## 24      28.5  14     M06   Male
## 25      22.0   8     M07   Male
## 26      22.0  10     M07   Male
## 27      24.5  12     M07   Male
## 28      26.5  14     M07   Male
## 29      24.0   8     M08   Male
## 30      21.5  10     M08   Male
## 31      24.5  12     M08   Male
## 32      25.5  14     M08   Male
## 33      23.0   8     M09   Male
## 34      20.5  10     M09   Male
## 35      31.0  12     M09   Male
## 36      26.0  14     M09   Male
```

Indexing and subsetting values within a data.frame

```
Orthodont$distance
```

```
##   [1] 26.0 25.0 29.0 31.0 21.5 22.5 23.0 26.5 23.0 22.5 24.0 27.5 25.5 27.5
##  [15] 26.5 27.0 20.0 23.5 22.5 26.0 24.5 25.5 27.0 28.5 22.0 22.0 24.5 26.5
##  [29] 24.0 21.5 24.5 25.5 23.0 20.5 31.0 26.0 27.5 28.0 31.0 31.5 23.0 23.0
##  [43] 23.5 25.0 21.5 23.5 24.0 28.0 17.0 24.5 26.0 29.5 22.5 25.5 25.5 26.0
##  [57] 23.0 24.5 26.0 30.0 22.0 21.5 23.5 25.0 21.0 20.0 21.5 23.0 21.0 21.5
##  [71] 24.0 25.5 20.5 24.0 24.5 26.0 23.5 24.5 25.0 26.5 21.5 23.0 22.5 23.5
##  [85] 20.0 21.0 21.0 22.5 21.5 22.5 23.0 25.0 23.0 23.0 23.5 24.0 20.0 21.0
##  [99] 22.0 21.5 16.5 19.0 19.0 19.5 24.5 25.0 28.0 28.0
```

```
head(Orthodont[1:5, 1:2])
```

```
##   distance age
## 1     26.0   8
## 2     25.0  10
## 3     29.0  12
## 4     31.0  14
## 5     21.5   8
```

```
Orthodont[, 1]
```

```
##   [1] 26.0 25.0 29.0 31.0 21.5 22.5 23.0 26.5 23.0 22.5 24.0 27.5 25.5 27.5
##  [15] 26.5 27.0 20.0 23.5 22.5 26.0 24.5 25.5 27.0 28.5 22.0 22.0 24.5 26.5
##  [29] 24.0 21.5 24.5 25.5 23.0 20.5 31.0 26.0 27.5 28.0 31.0 31.5 23.0 23.0
##  [43] 23.5 25.0 21.5 23.5 24.0 28.0 17.0 24.5 26.0 29.5 22.5 25.5 25.5 26.0
##  [57] 23.0 24.5 26.0 30.0 22.0 21.5 23.5 25.0 21.0 20.0 21.5 23.0 21.0 21.5
##  [71] 24.0 25.5 20.5 24.0 24.5 26.0 23.5 24.5 25.0 26.5 21.5 23.0 22.5 23.5
##  [85] 20.0 21.0 21.0 22.5 21.5 22.5 23.0 25.0 23.0 23.0 23.5 24.0 20.0 21.0
##  [99] 22.0 21.5 16.5 19.0 19.0 19.5 24.5 25.0 28.0 28.0
```

```
OrthoF <- subset(Orthodont, Sex == "Female")
is.data.frame(OrthoF)
```

```
## [1] TRUE
```

텍스트 파일 변환 없이 R data.frame을 그래도 저장하고 부를 수 있다.

```
nrow(OrthoF); ncol(OrthoF)

## [1] 44
## [1] 4

save(OrthoF, file = "OrthoF.RData")
#
#
#
load("OrthoF.Rdata")
#
#
#
```

Universal text-format of table data

- CSV (*.csv): Comma-separated values
- TSV (*.tsv *.txt): Tab-separated values

```
CSV
"distance","age","Subject","Sex"
26,8,"M01","Male"
25,10,"M01","Male"
29,12,"M01","Male"
31,14,"M01","Male"
21.5,8,"M02","Male"
.
.
.
```

```
TSV
"distance"  "age"  "Subject"  "Sex"
26  8  "M01"  "Male"
25  10  "M01"  "Male"
29  12  "M01"  "Male"
31  14  "M01"  "Male"
21.5  8  "M02"  "Male"
.
.
.
```

Try:

```
write.table(Orthodont, "Orthodont.csv", sep = ",", row.names = FALSE)
```

Compare:

```
write.table(Orthodont, "Orthodont.tsv", sep = "\t", row.names = FALSE)
```

```r
orthoFromCSV <- read.table("Orthodont.csv", sep = ",", header = TRUE)
is.data.frame(orthoFromCSV)
```

```
## [1] TRUE
```

```r
str(orthoFromCSV)
```

```
## 'data.frame': 108 obs. of  4 variables:
##  $ distance: num  26 25 29 31 21.5 22.5 23 26.5 23 22.5 ...
##  $ age     : int  8 10 12 14 8 10 12 14 8 10 ...
##  $ Subject : Factor w/ 27 levels "F01","F02","F03",..: 12 12 12 12 13 13 13 13 14 14 ...
##  $ Sex     : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
```

```r
orthoFromCSV2 <- read.table("Orthodont.csv", sep = ",",
    stringsAsFactors  = FALSE, header = TRUE)
str(orthoFromCSV2)
```

```
## 'data.frame': 108 obs. of  4 variables:
##  $ distance: num  26 25 29 31 21.5 22.5 23 26.5 23 22.5 ...
##  $ age     : int  8 10 12 14 8 10 12 14 8 10 ...
##  $ Subject : chr  "M01" "M01" "M01" "M01" ...
##  $ Sex     : chr  "Male" "Male" "Male" "Male" ...
```

Reading CSV from Distant Source

```
# https://vincentarelbundock.github.io/Rdatasets/csv/boot/calcium.csv

calcium <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/boot/calcium.csv",
header = TRUE)
str(calcium)

## 'data.frame': 27 obs. of  3 variables:
##  $ X   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ time: num  0.45 0.45 0.45 1.3 1.3 1.3 2.4 2.4 2.4 4 ...
##  $ cal : num  0.3417 -0.00438 0.82531 1.77967 0.95384 ...

head(calcium)

##   X time      cal
## 1 1 0.45  0.34170
## 2 2 0.45 -0.00438
## 3 3 0.45  0.82531
## 4 4 1.30  1.77967
## 5 5 1.30  0.95384
## 6 6 1.30  0.64080
```

Orthodont 자료철에서 연령별, 성별로 구분하여 길이(**distance**)의 요약통계량을[a] 구하라.

[a] 평균은 mean(), quantile(), summary() 함수 이용

Orthodont 자료철에서 연령별, 성별로 구분하여 길이(**distance**)의 요약통계량을[a] 구하라.

[a] 평균은 mean(), quantile(), summary() 함수 이용

```
mean(Orthodont$distance)
```

```
## [1] 24.02315
```

```
quantile(Orthodont$distance)
```

```
##    0%   25%   50%   75%  100%
## 16.50 22.00 23.75 26.00 31.50
```

```
summary(Orthodont$distance)
```

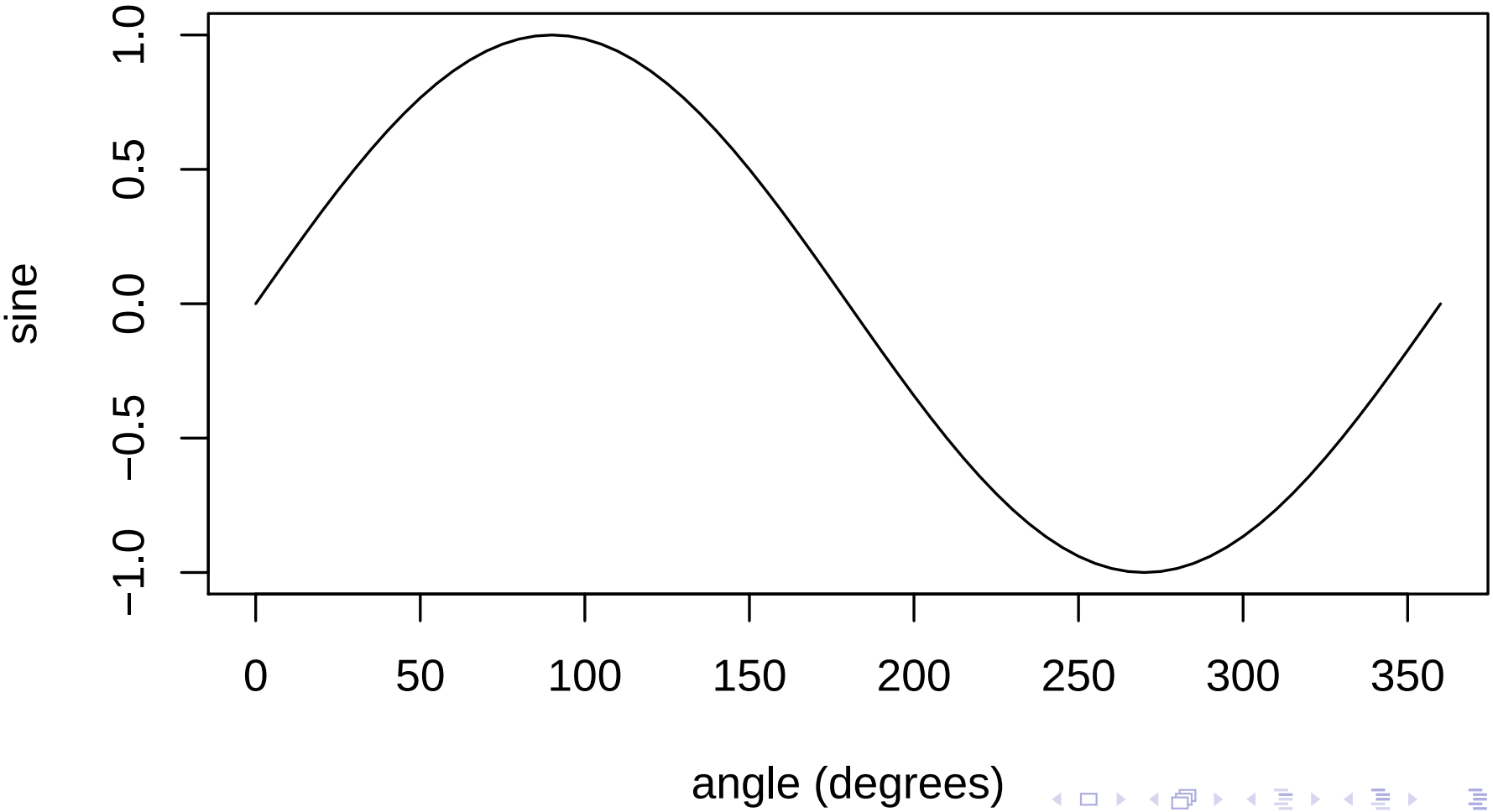
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.50   22.00   23.75   24.02   26.00   31.50
```

Notice

Excel data (*.xlsx), a famous spreadsheet format, SPSS data (*.dat), or SAS data (*. bdat) can be easily imported to R, which is not covered in this course of lecture.

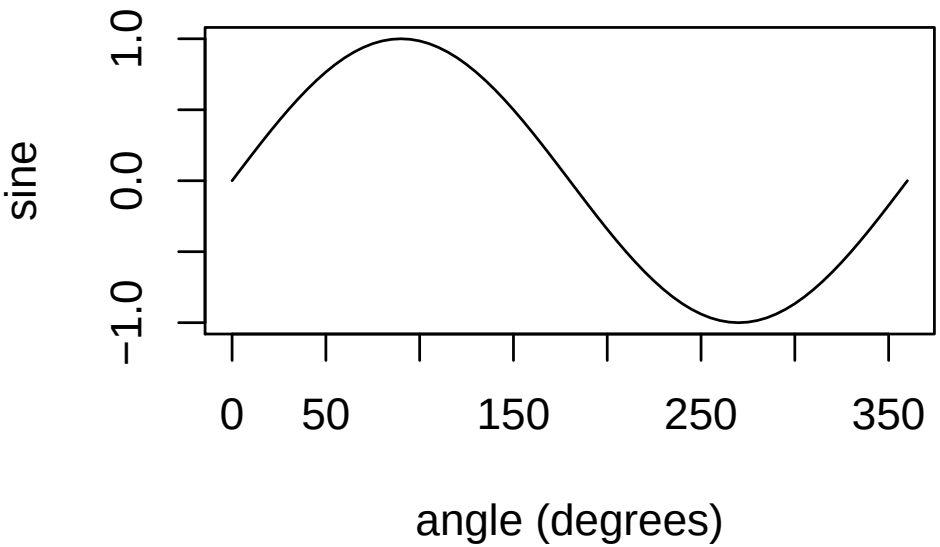
Graphing

A Typical Scientific Graphic Chart



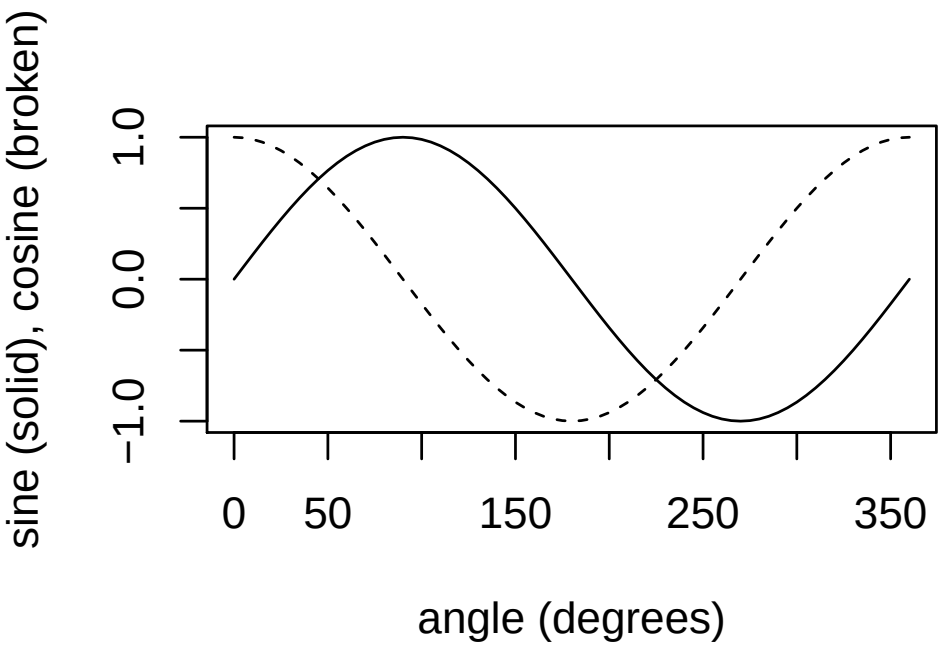
Grammar of Graphics

To understand before graphing a chart, we always split a chart into *DATA*, *ELEMENT* and *GUIDE*.



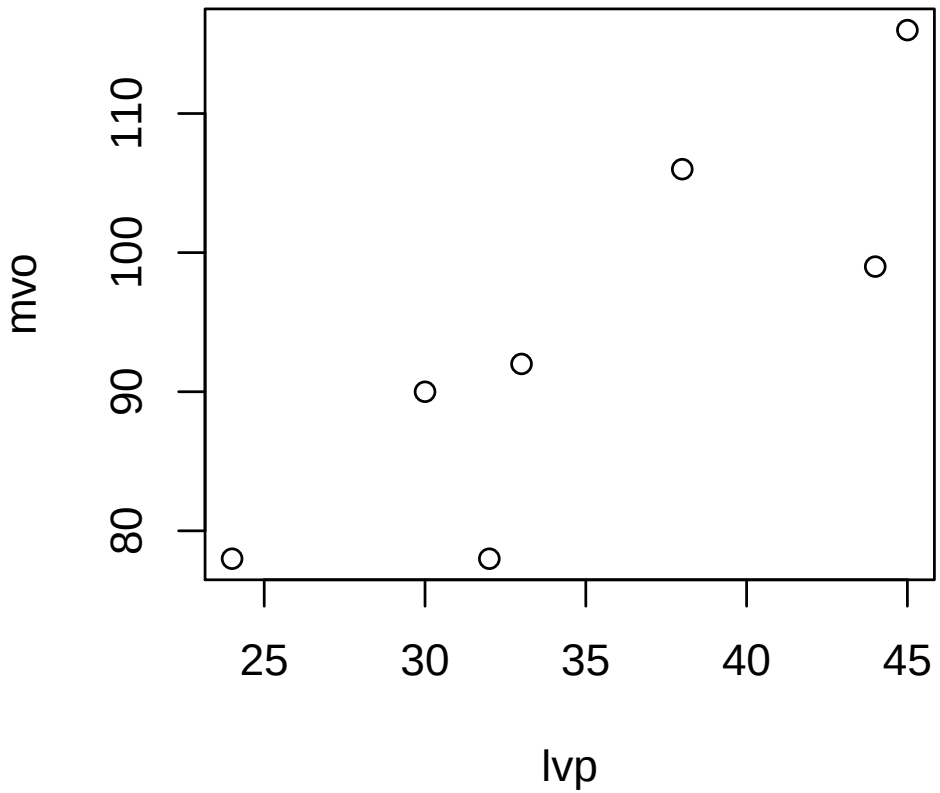
Component	Content
ELEMENT	line (position (angle * its sine value))
GUIDE	axis (dim(1), limit(0, 350), tick(50), label = "angle (degrees)")
GUIDE	axis (dim(2), limit(-1.0, +1.0), tick(0.5), label = "sine ")

Grammar of Graphics



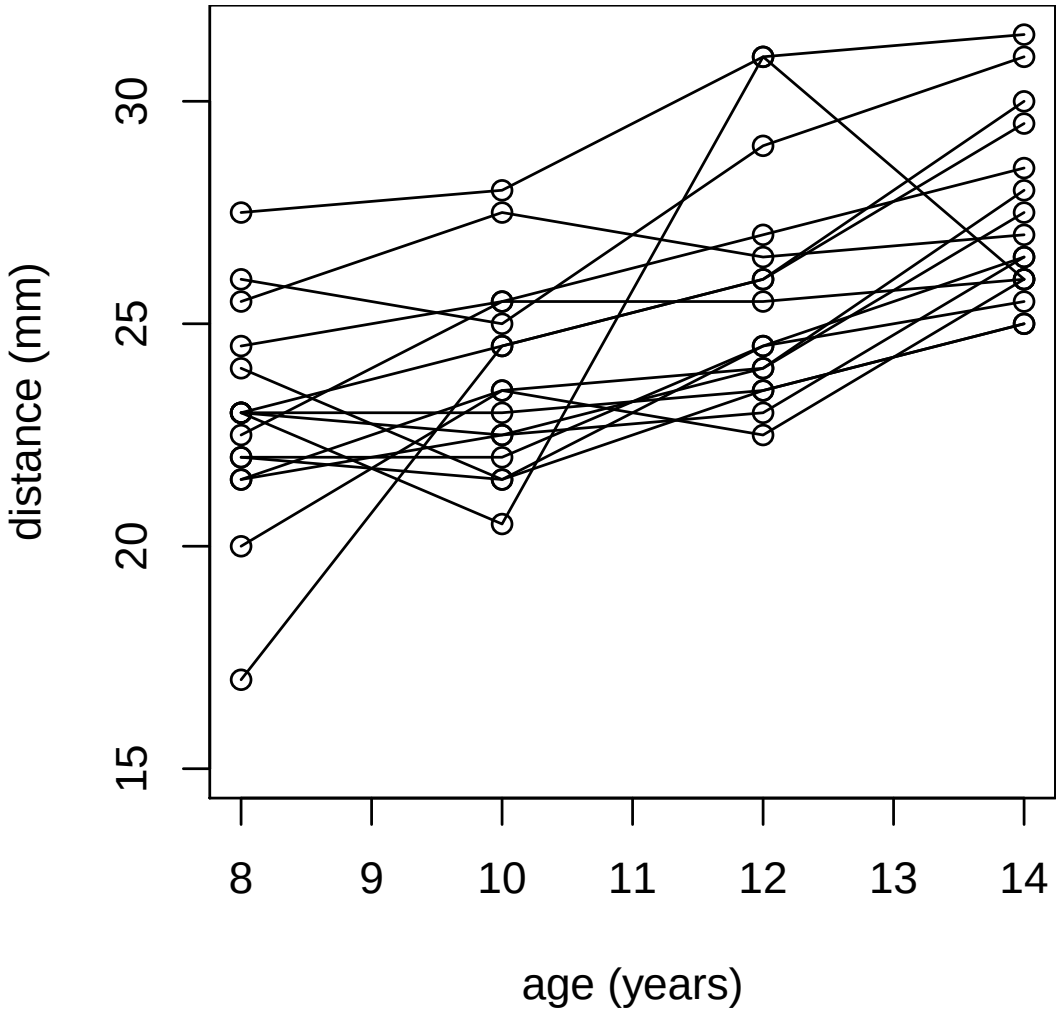
Component	Content
ELEMENT	line (position (angle * its sine value), linetype = "solid")
ELEMENT	line (position (angle * its cossine value), linetype = "broken")
GUIDE	axis (dim(1), limit(0, 350), tick(50), label = "angle (degrees)")
GUIDE	axis (dim(2), limit(-1.0, +1.0), tick(0.5), label = "sine (solid), cosine (broken)")

Grammar of Graphics



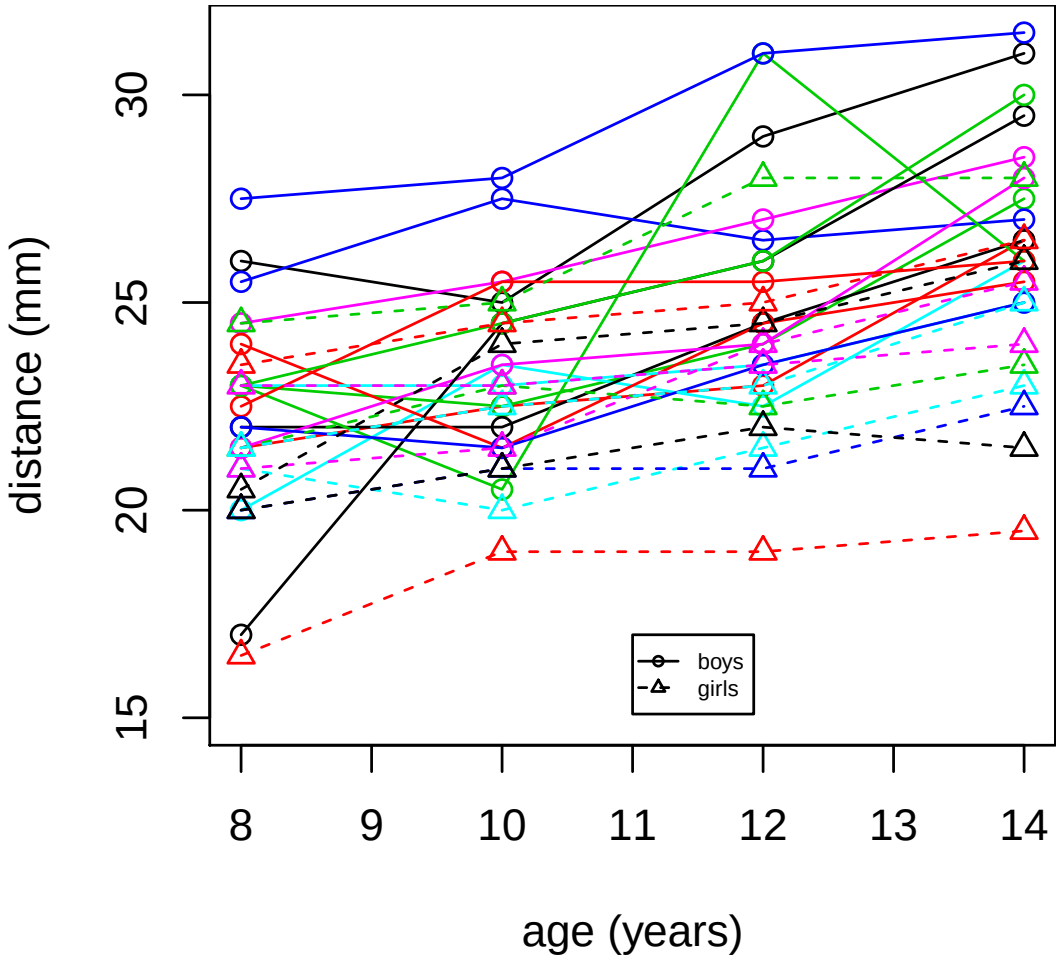
Component	Content
ELEMENT	point (position (lvp * mvo))
GUIDE	axis (dim(1), limit(23, 45), tick(5), label = "lvp")
GUIDE	axis (dim(2), limit(77, 117), tick(10), label = "mvo")

Grammar of Graphics



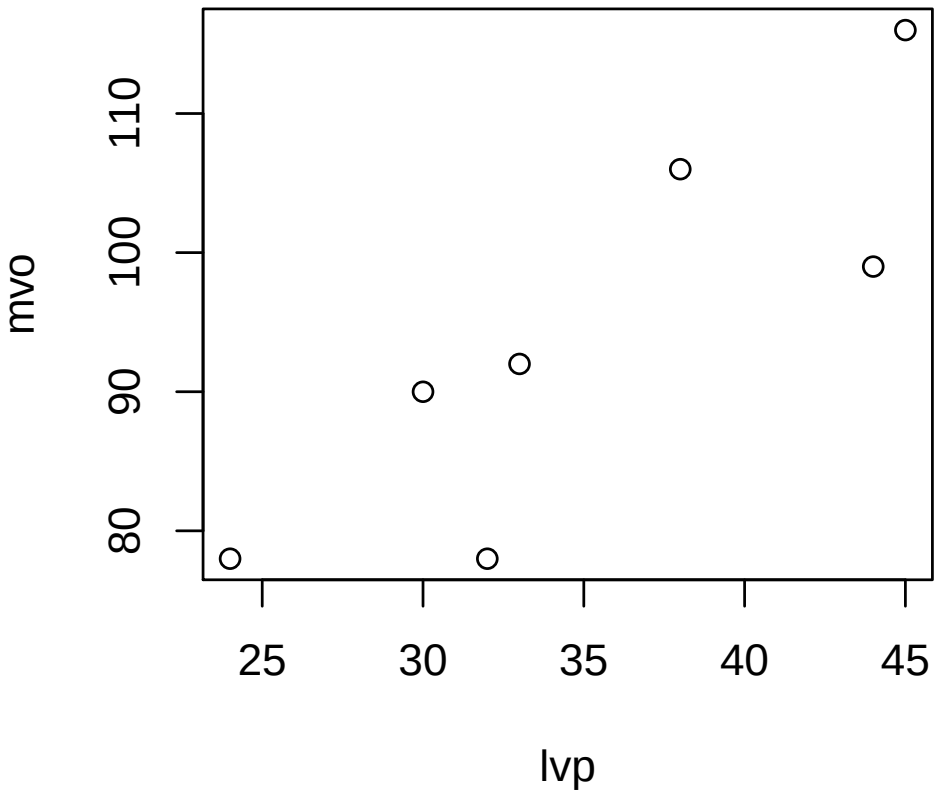
Component	Content
ELEMENT	point (position (age * distance))
ELEMENT	line (position (age * distance))
GUIDE	axis (dim(1), discrete(8, 10, 12, 14), tick(1), label = "age (years)")
GUIDE	axis (dim(2), limit(15, 32), tick(5), label = "distance (mm)")

Grammar of Graphics



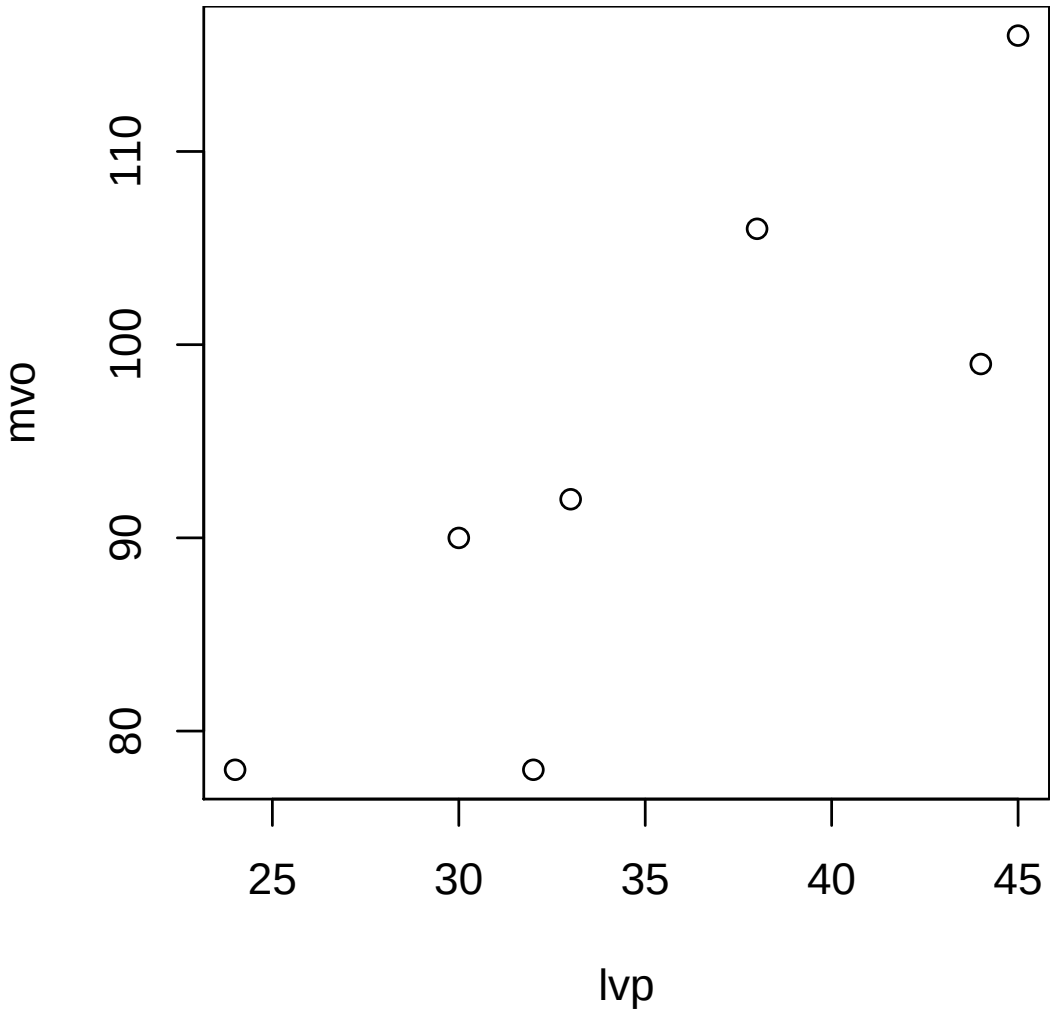
Component	Content
ELEMENT	point (position (age * distance), shape (gender), color (subject))
ELEMENT	line (position (age * distance), shape (gender), color (subjec))
GUIDE	legend.point (dim(1), label(gender))
GUIDE	legend.line (dim(1), label(gender))
GUIDE	axis (dim(1), discrete(8, 10, 12, 14), tick(1), label = "age (years)")
GUIDE	axis (dim(2), limit(15, 32), tick(5), label = "distance (mm)")

Grammar of Graphics



Component	Content
ELEMENT	point (position (lvp * mvo))
GUIDE	axis (dim(1), limit(23, 45), tick(5), label = "lvp")
GUIDE	axis (dim(2), limit(77, 117), tick(10), label = "mvo")

Grammar of Graphics



Component	Content
ELEMENT	point (position (lvp * mvo))
GUIDE	axis (dim(1), limit(23, 45), tick(5), label = "lvp")
GUIDE	axis (dim(2), limit(77, 117), tick(10), label = "mvo")

Component	Content	
ELEMENT	point (position (lvp * mvo))	plot(lvp, mvo, data = dogs)
GUIDE	axis (dim(1), limit(23, 45), tick(5), label = "lvp")	xlab = "lvp", xlim = c(23, 45)
GUIDE	axis (dim(2), limit(77, 117), tick(10), label = "mvo")	ylab = "mvo", ylim = c(77, 117)

Try:

```
require(boot)
data(dogs, package = "boot")
plot(mvo ~ lvp, data = dogs,
    xlab = "lvp", xlim = c(23, 45),
        ylab = "mvo", ylim = c(77, 117))
```

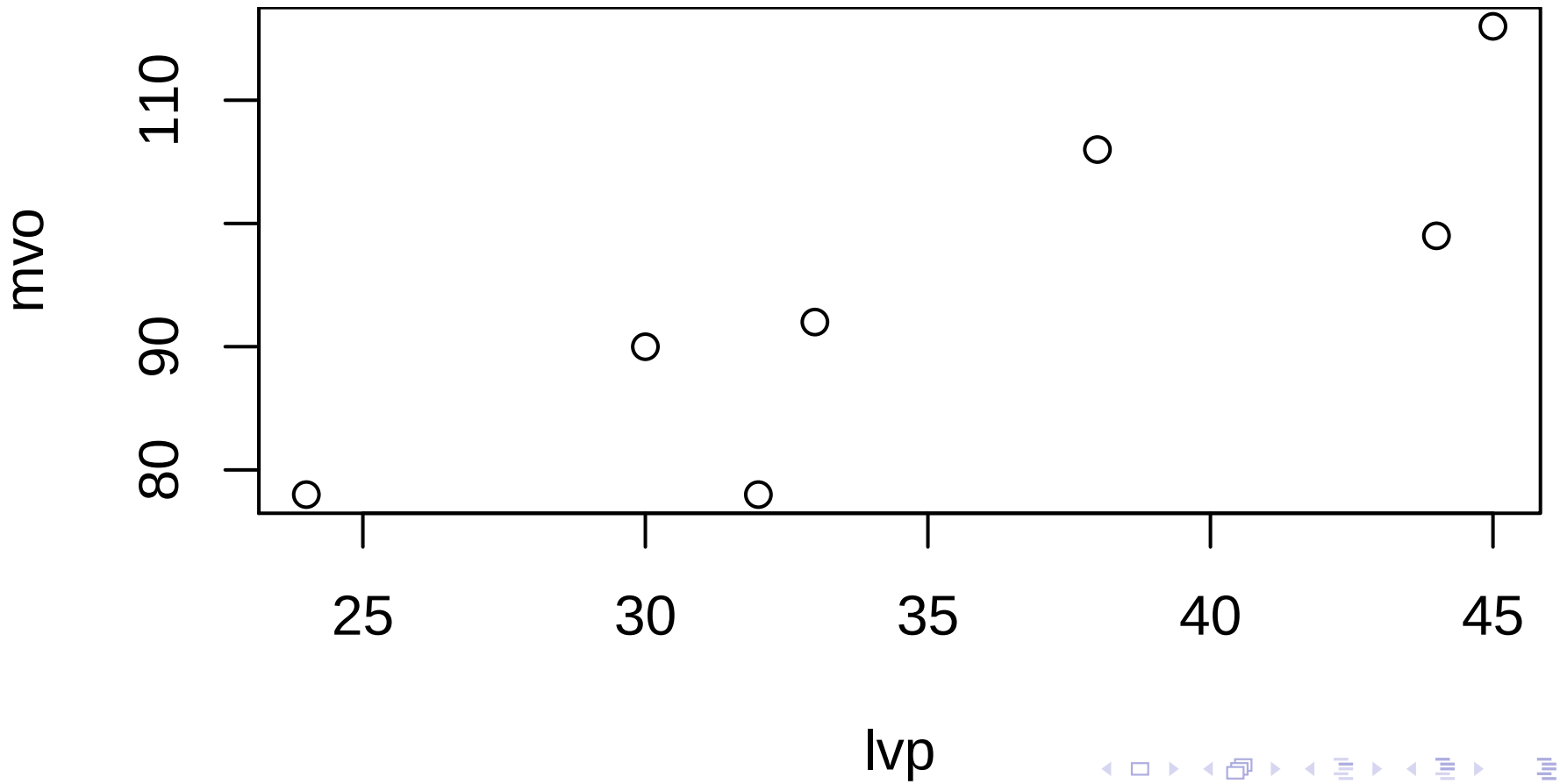
Compare:

```
plot(mvo ~ lvp, data = dogs)
plot(dogs$lvp, dogs$mvo)
```

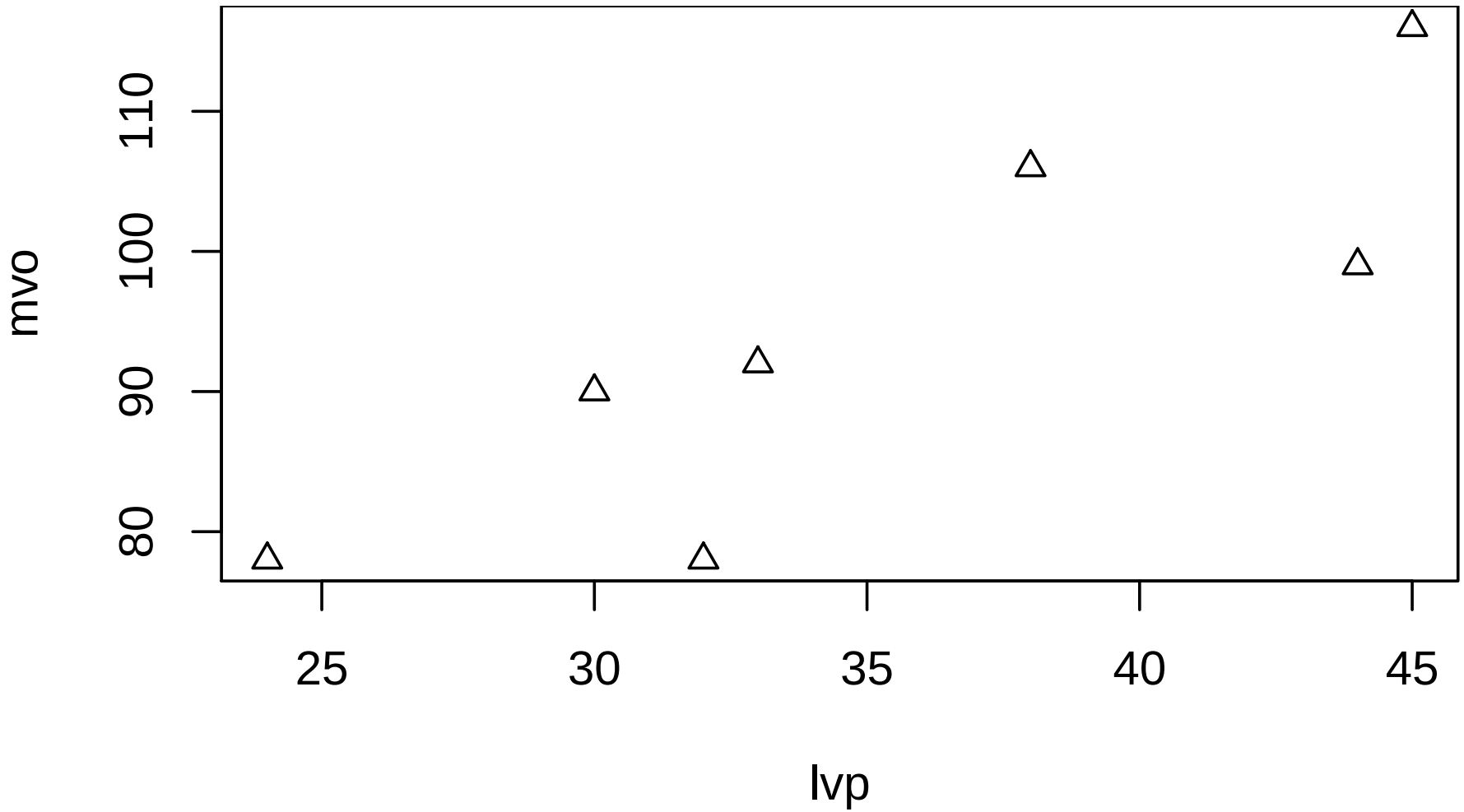
```r
t(dogs)
```

```
##       1  2   3  4   5  6  7
## mvo  78 92 116 90 106 78 99
## lvp  32 33  45 30  38 24 44
```

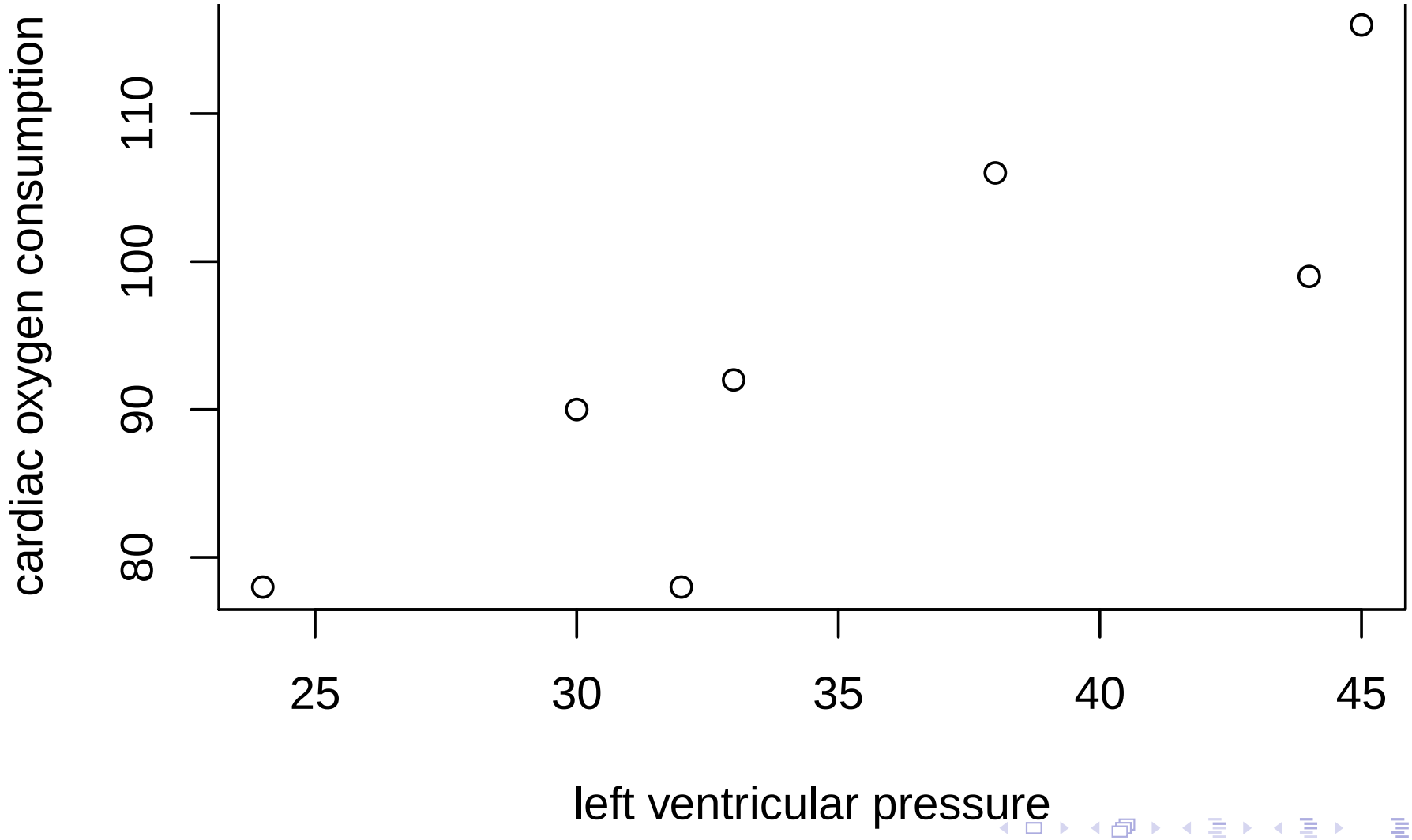
```r
par(mar = c(5, 4, 0, 2))
plot(mvo ~ lvp, data = dogs)
```



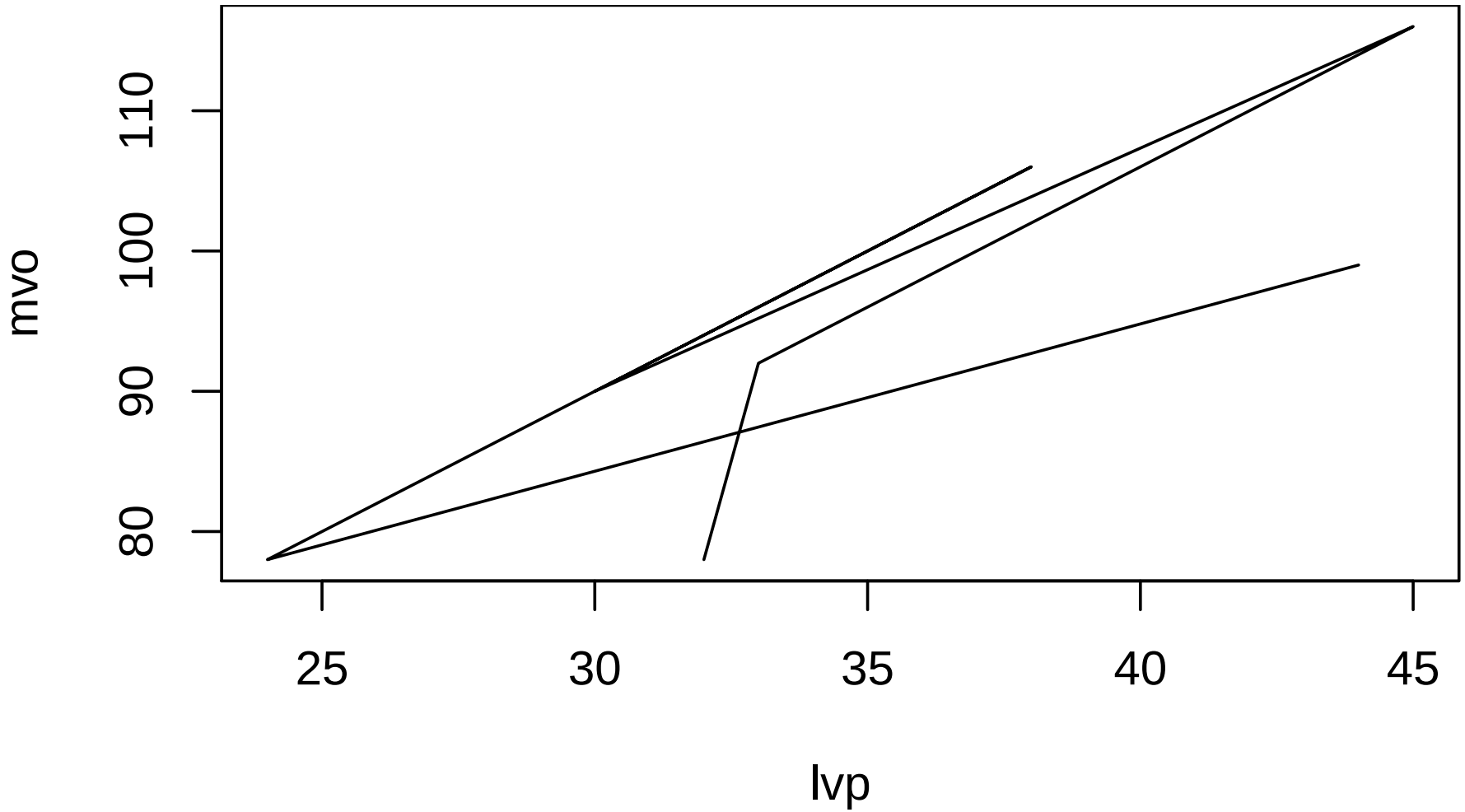
```r
par(mar = c(5, 4, 0, 2))
plot(mvo ~ lvp, data = dogs,
     pch = 2)
```



```
par(mar = c(5, 4, 0, 2))
plot(mvo ~ lvp, data = dogs,
    xlab = "left ventricular pressure",
        ylab = "cardiac oxygen consumption")
```



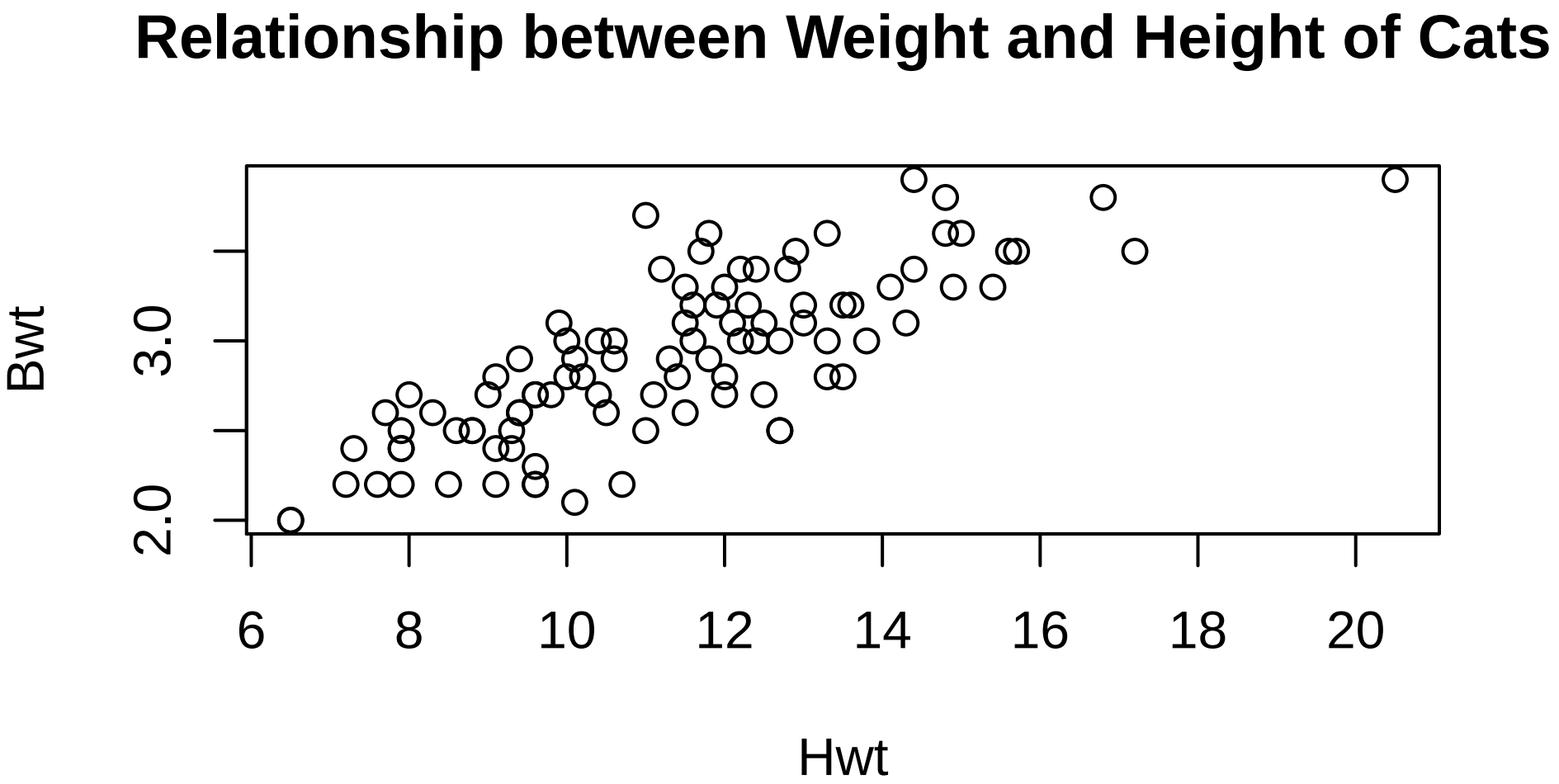
```
par(mar = c(5, 4, 0, 2))
plot(mvo ~ lvp, data = dogs,
    type = "l")
```



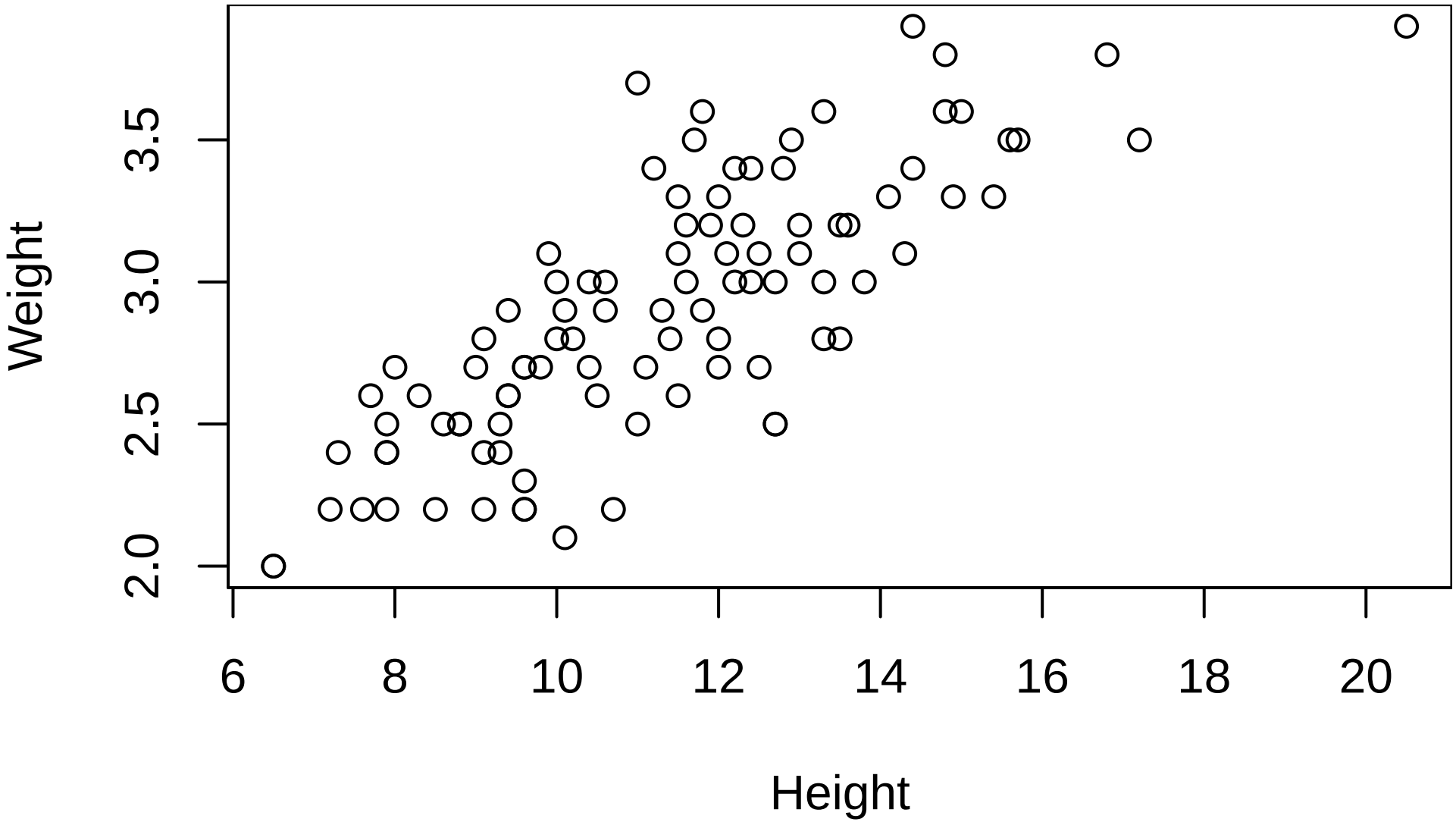
97마리의 고양이의 체중과 신장 자료로부터 체중과 신장의 기초통계량을 구한 뒤 둘 사이의 관계를 산포도로 표현한 뒤 "Relationship between Weight and Height of Cats"로 붙이시오. (자료는 https://vincentarelbundock.github.io/Rdatasets/csv/boot/catsM.csv에 있다.)

```
catsM <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/boot/catsM.csv", header =
str(catsM)

## 'data.frame': 97 obs. of  4 variables:
##  $ X  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sex: Factor w/ 1 level "M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Bwt: num  2 2 2.1 2.2 2.2 2.2 2.2 2.2 2.2 2.2 ...
##  $ Hwt: num  6.5 6.5 10.1 7.2 7.6 7.9 8.5 9.1 9.6 9.6 ...

head(catsM)

##   X Sex Bwt  Hwt
## 1 1   M 2.0  6.5
## 2 2   M 2.0  6.5
## 3 3   M 2.1 10.1
## 4 4   M 2.2  7.2
## 5 5   M 2.2  7.6
## 6 6   M 2.2  7.9

summary(catsM)

##        X      Sex         Bwt             Hwt
##  Min.   : 1   M:97   Min.   :2.0   Min.   : 6.50
##  1st Qu.:25          1st Qu.:2.5   1st Qu.: 9.40
##  Median :49          Median :2.9   Median :11.40
##  Mean   :49          Mean   :2.9   Mean   :11.32
##  3rd Qu.:73          3rd Qu.:3.2   3rd Qu.:12.80
##  Max.   :97          Max.   :3.9   Max.   :20.50
```

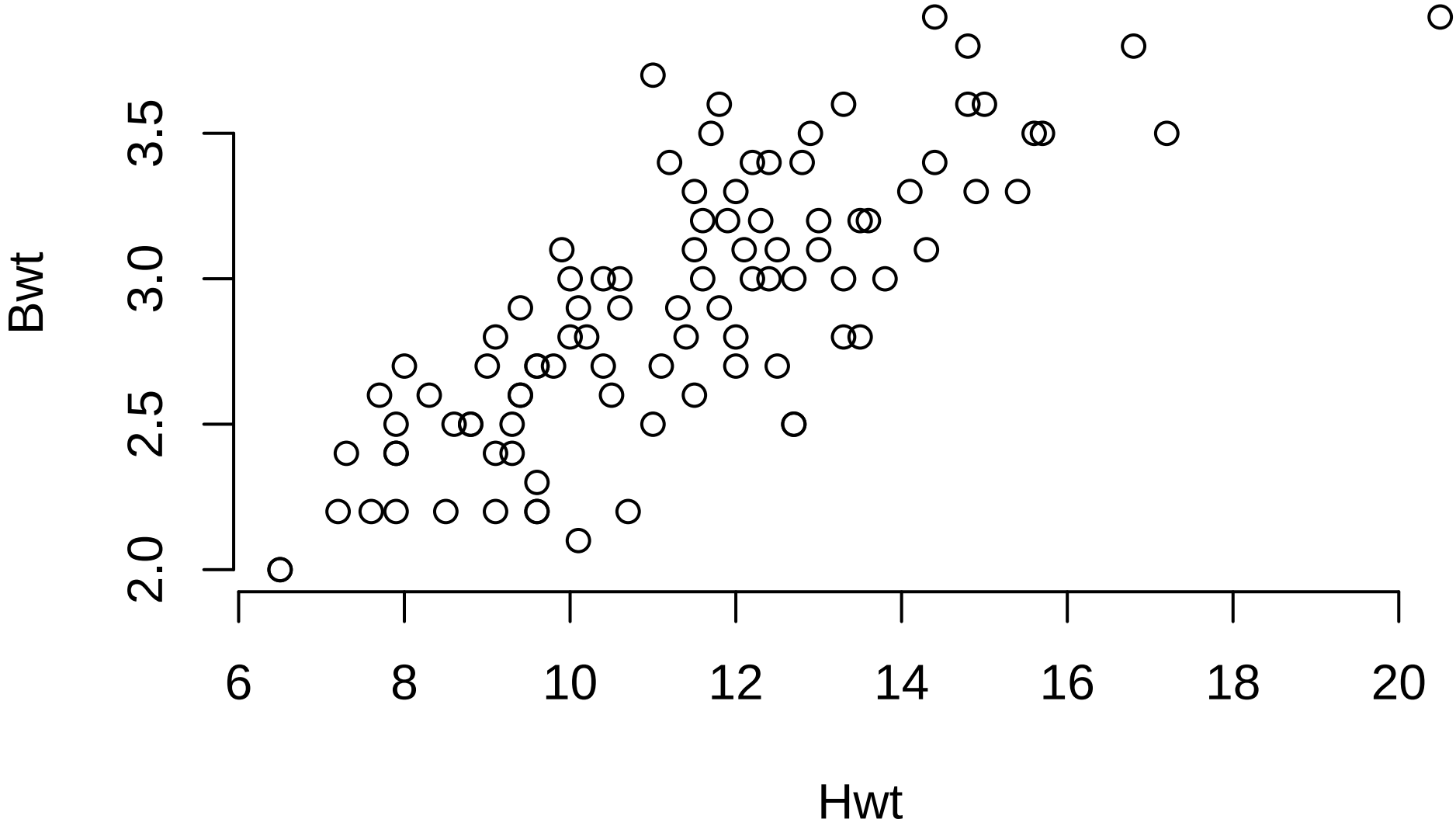
```
plot(Bwt ~ Hwt, data = catsM,
    main = "Relationship between Weight and Height of Cats")
```



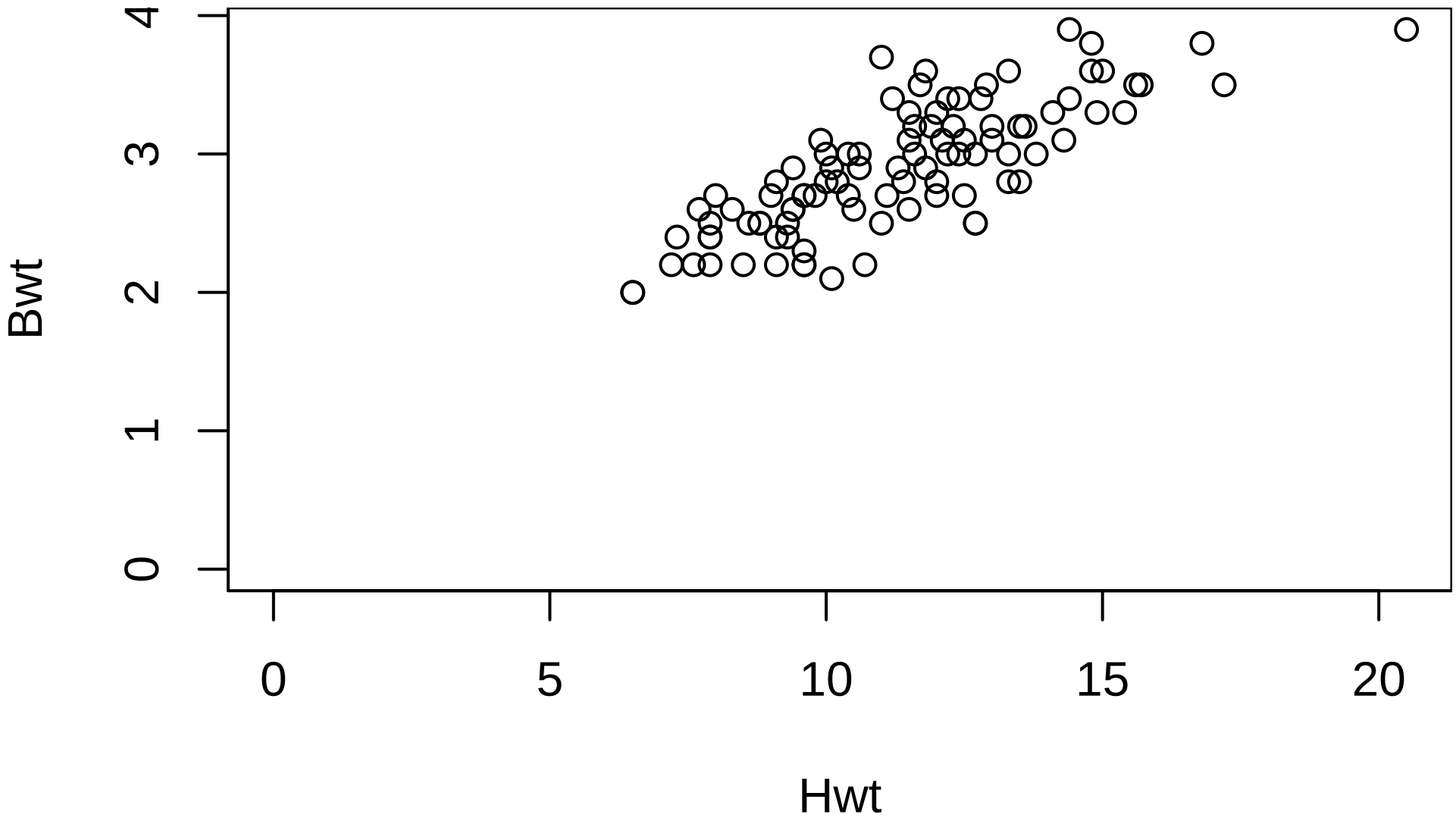
고양이 체중신장 그래프에서 가로축과 세로축의 이름에 변수명이 아닌 설명을 넣으려면?



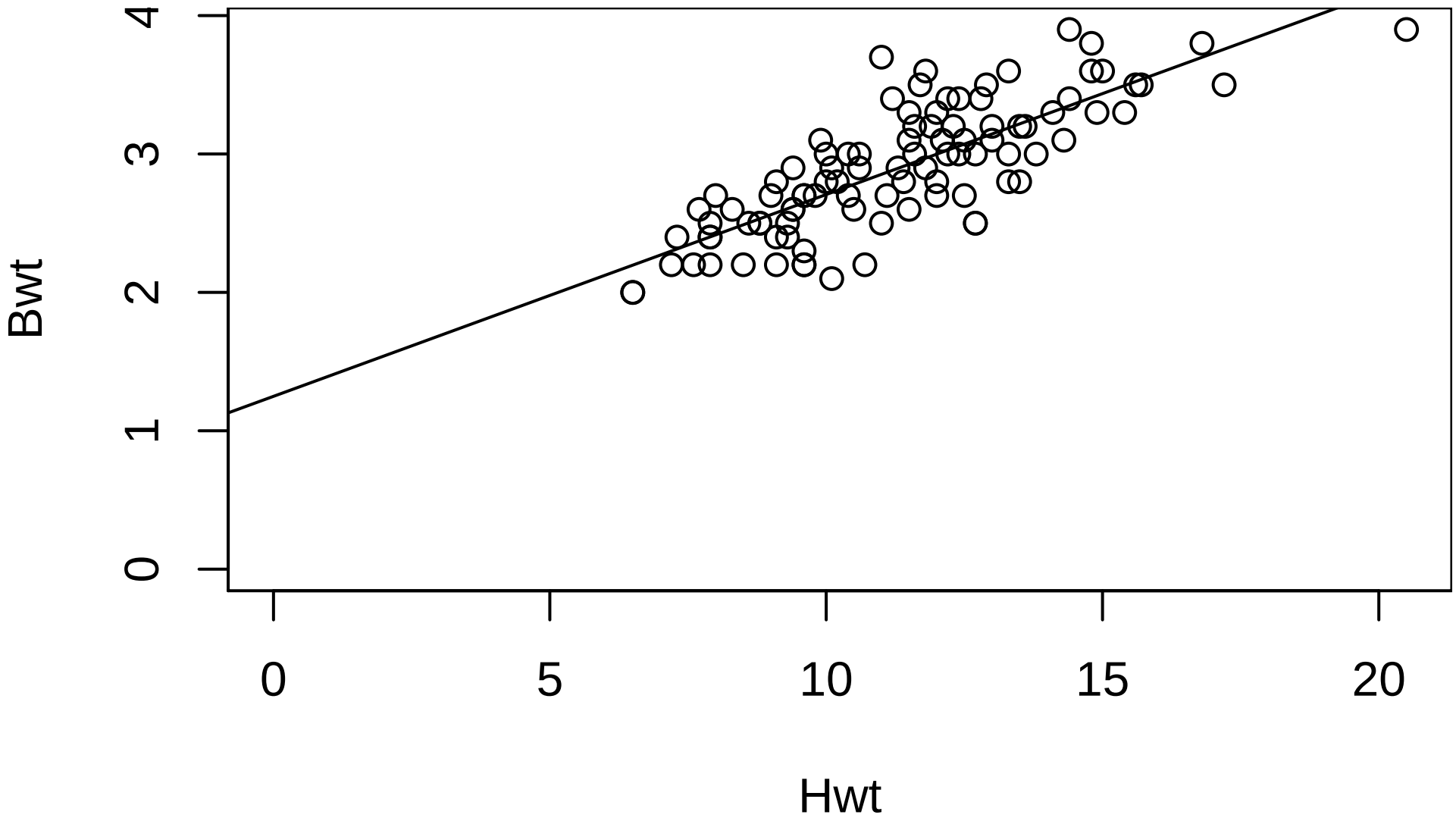
고양이 체중신장 그래프에서 가로축과 세로축을 하나씩만 남기려면?



축의 원점을 0, 0으로 잡으려면?



회귀직선을 넣을 수 있을까?



Matrix plot with **matplot()**

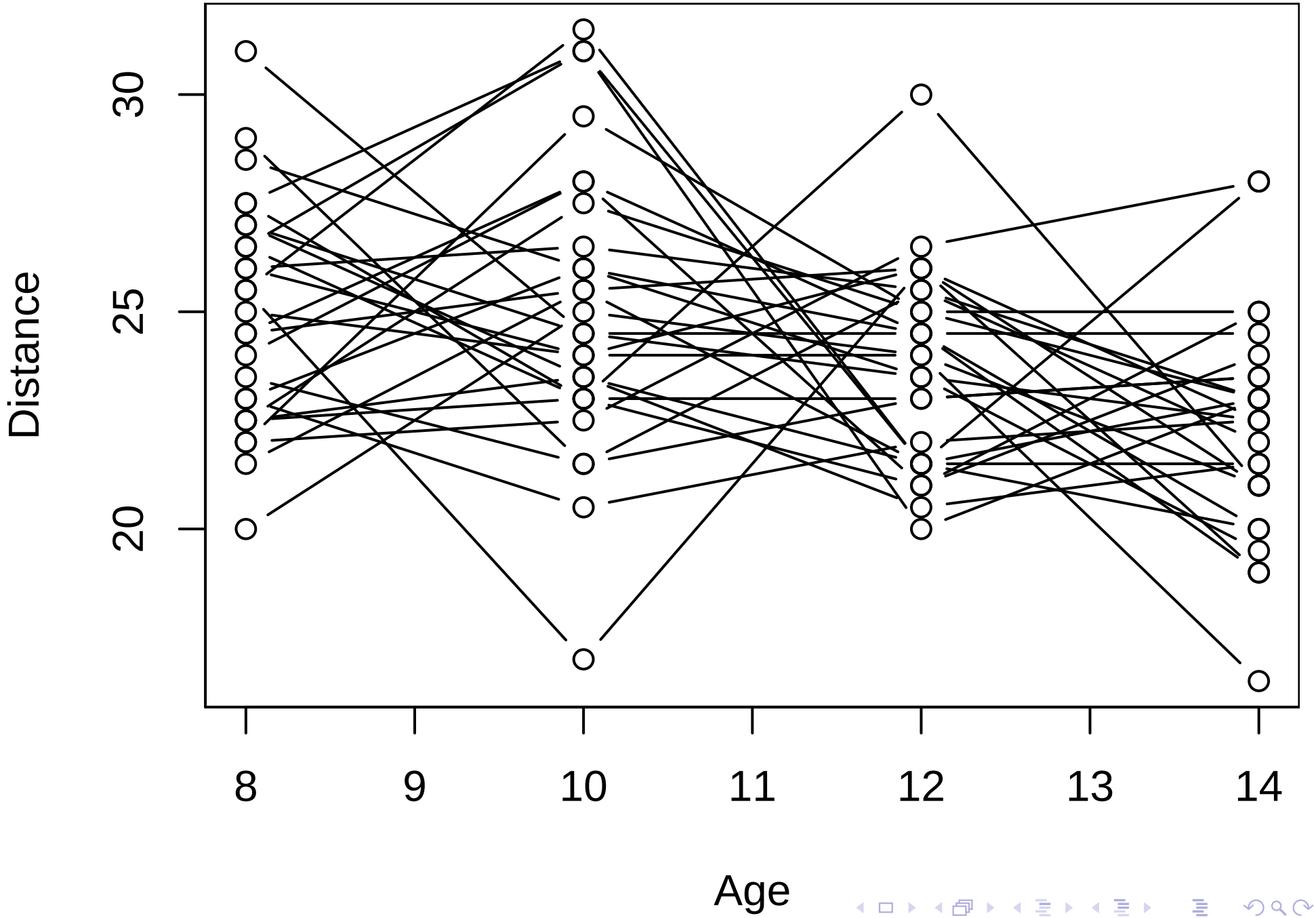
```
head(Orthodont)
```

```
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

```
distanceAsMatrix <- matrix(Orthodont$distance, byrow = TRUE, nrow = 4)
table(Orthodont$Sex) / 4
```

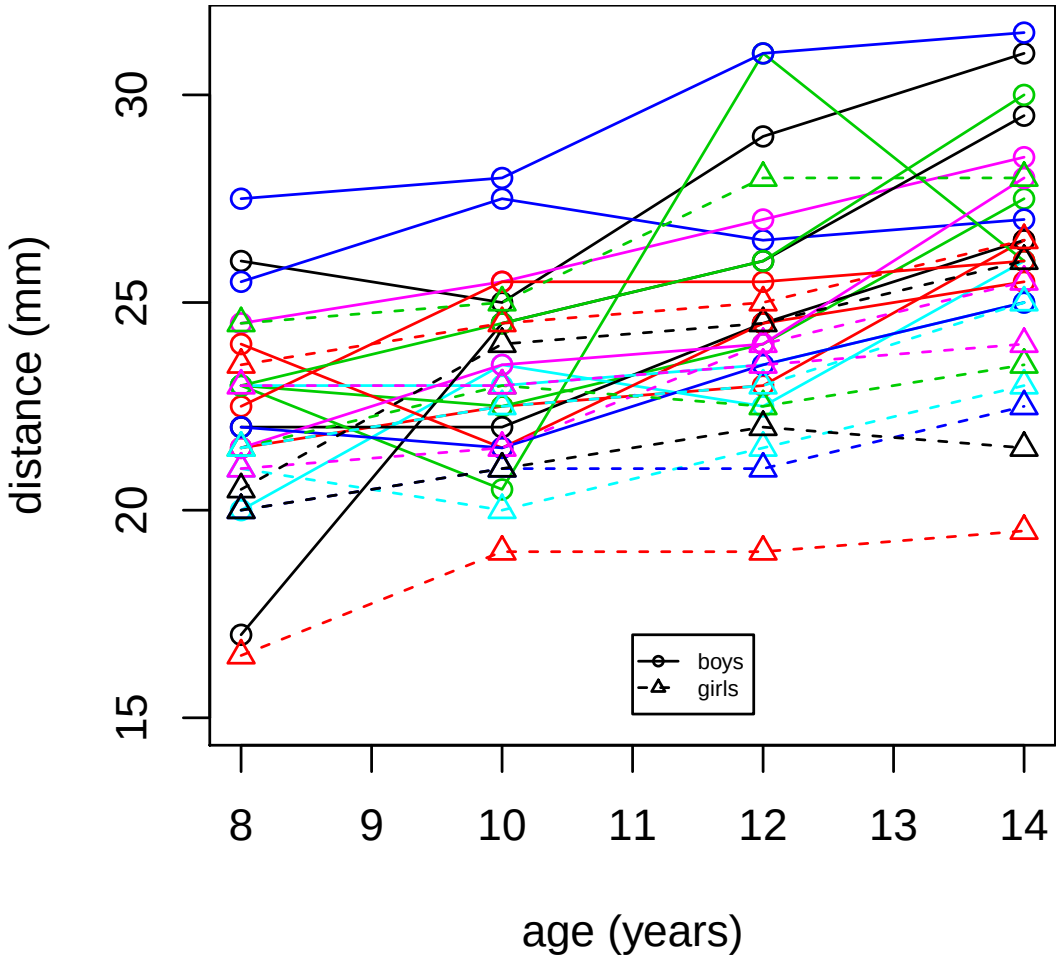
```
##
##   Male Female
##     16     11
```

```
matplot(c(8, 10, 12, 14), distanceAsMatrix, type = "b",
    pch = 1, col = 1, lty = 1,
        xlab = "Age",
        ylab = "Distance")
```


Distance
20
25
30
8
9
10
11
12
13
14
Age

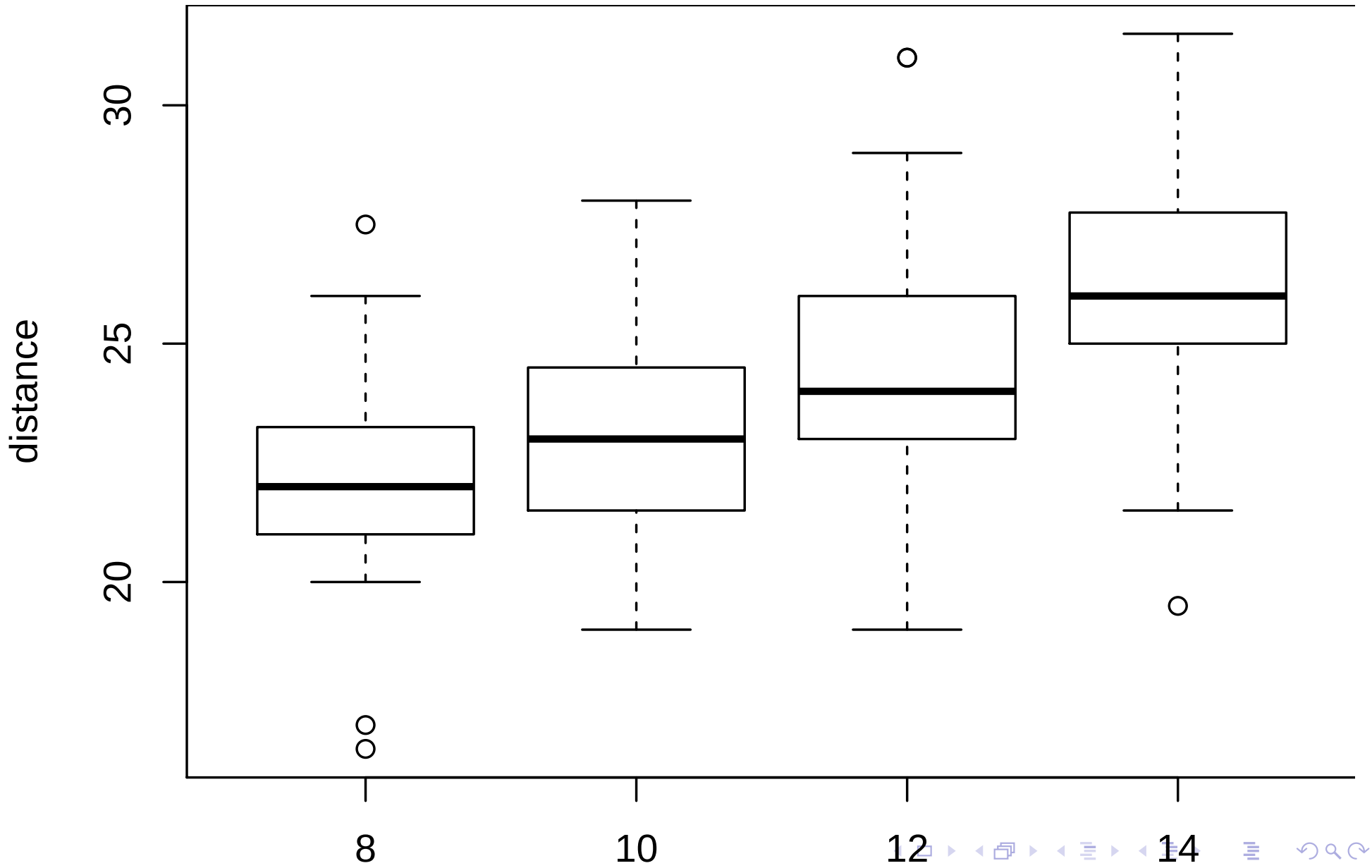
앞의 Orthodont 자료를 matplot()으로 연령별 변화를 도식하되 남자아이와 여자아이를 구별하는 방법이 있을까?

Grammar of Graphics



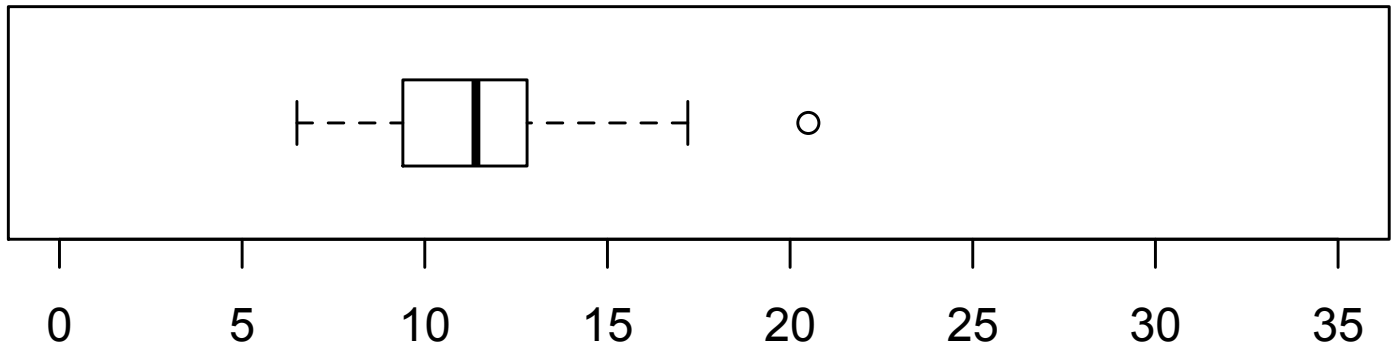
Component	Content
ELEMENT	point (position (age * distance), shape (gender), color (subject))
ELEMENT	line (position (age * distance), shape (gender), color (subjec))
GUIDE	legend.point (dim(1), label(gender))
GUIDE	legend.line (dim(1), label(gender))
GUIDE	axis (dim(1), discrete(8, 10, 12, 14), tick(1), label = "age (years)")
GUIDE	axis (dim(2), limit(15, 32), tick(5), label = "distance (mm)")

boxplot: 수치형 자료의 고급 통계도식법



Schema as an element

Component	Content
ELEMENT	schema (position (bin.quantile.letter(distance * age)))
GUIDE	axis (dim(1), discrete(8, 10, 12, 14), label = "age")
GUIDE	axis (dim(2), limit(15, 32), tick(5), label = "distance")




```
summary(catsM$Hwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.50    9.40   11.40   11.32   12.80   20.50
```

```
median(catsM$Hwt) - IQR(catsM$Hwt) * 1.5
```

```
## [1] 6.3
```

```
median(catsM$Hwt) + IQR(catsM$Hwt) * 1.5
```

```
## [1] 16.5
```

Orthodont 자료철을 이용해서 다음의 element와 guide를 가지는 그래프를 완성하라.

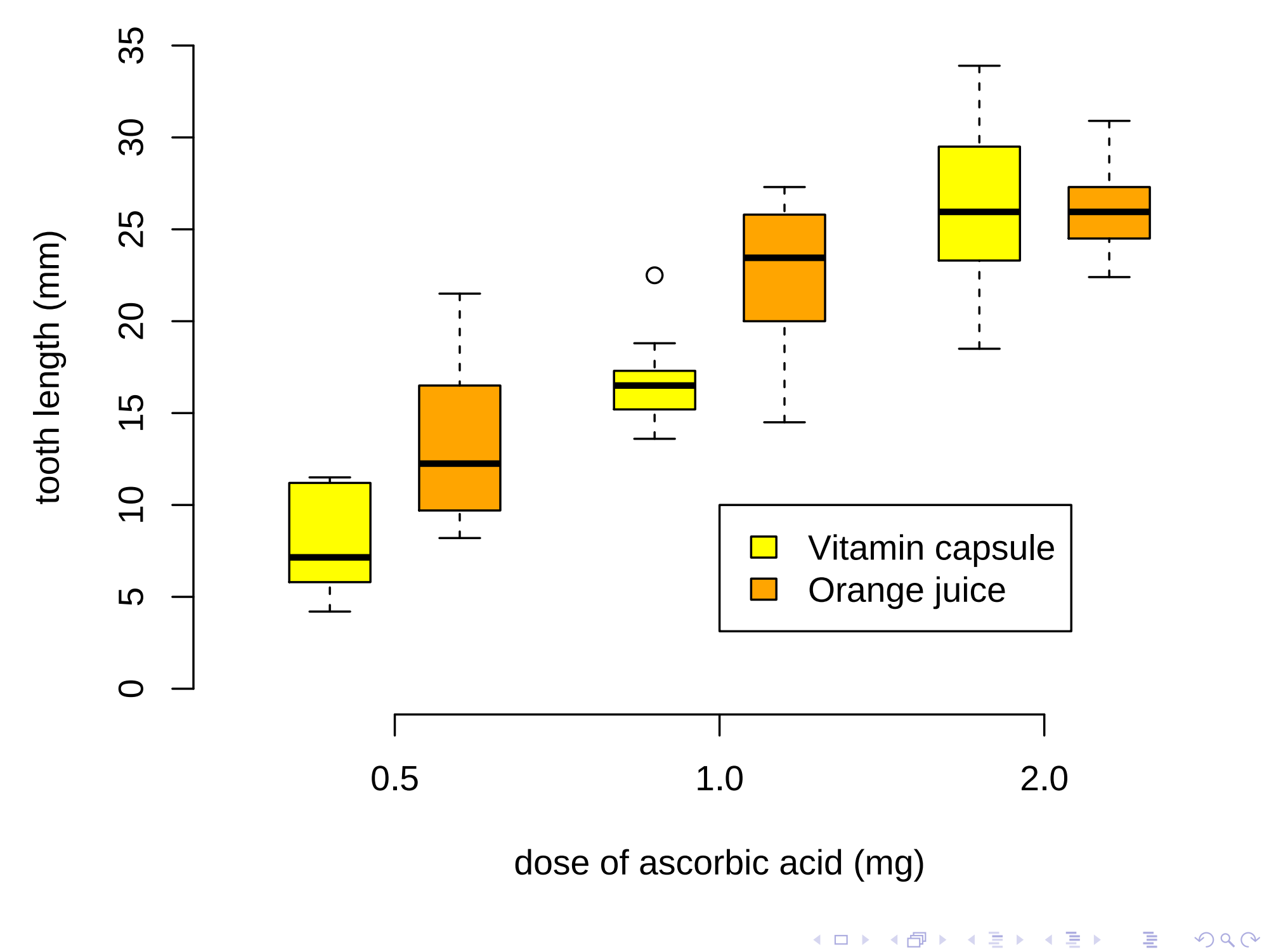
Component	Content
ELEMENT	schema (position (bin.quantile.letter(distance * (age * Sex)), color (Sex)))
GUIDE	axis (dim(1), discrete(8, 10, 12, 14), label = "age")
GUIDE	axis (dim(2), limit(15, 32), tick(5), label = "distance")

Crampton은 음식물 내 아스코르브산이 용량 별로 실험동물의 치아 성장에 미치는 효과를 알고자 총 60마리의 기니피그를 골라 비타민캡슐과 오렌지주스로 나누어 실험하였다[a] R의 내장 자료철 **ToothGrowth**가 이 실험자료를 가지고 있으며 다음처럼 변수 **len** (numeric. 치아 길이), **supp** (factor. 오렌지주스는 OJ, 비타민캡슐은 VC로 코딩), **dose** (numeric. 용량)로 구성되어 있다:

```
str(ToothGrowth)

## 'data.frame': 60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

[a] Crampton, E. W. (1947). The growth of the odontoblast of the incisor teeth as a criterion of vitamin C intake of the guinea pig. The Journal of Nutrition, 33(5), 491–504. doi: 10.1093/jn/33.5.491.

이 **ToothGrowth** 자료에서 치아의 성장을 아스코르브산 용량에 따라 도해하면서 오렌지주스와 비타민캡슐을 색깔로 대별할 수 있도록 하되, R 기본함수인 **boxplot()**을 이용하라.


tooth length (mm)
0
5
10
15
20
25
30
35
0.5
1.0
2.0
dose of ascorbic acid (mg)
Vitamin capsule
Orange juice

- Read more about “Grammar of Graphics” from *The Grammar of Graphics (2nd Edition. 2005)* by Leland Wilkinson
- Look around more R graphics by googling an “R graphic gallery.”