

## Appendix C

# Quick introduction to R

### C.1 Reading the data in

The first step is to read the data in. You can use the `read.table()` or `scan()` function to read data in from outside R. You can also use the `data()` function to access data already available within R.

```
> data(stackloss)
> stackloss
  Air.Flow Water.Temp Acid.Conc. stack.loss
1      80      27      89      42
2      80      27      88      37
... stuff deleted ...
21     70      20      91      15
```

Type

```
> help(stackloss)
```

We can check the dimension of the data:

```
> dim(stackloss)
[1] 21  4
```

### C.2 Numerical Summaries

One easy way to get the basic numerical summaries is:

```
> summary(stackloss)
  Air.Flow      Water.Temp      Acid.Conc.      stack.loss
Min.   :50.0  Min.   :17.0  Min.   :72.0  Min.    : 7.0
1st Qu.:56.0  1st Qu.:18.0  1st Qu.:82.0  1st Qu.:11.0
Median :58.0  Median :20.0  Median :87.0  Median :15.0
Mean   :60.4  Mean   :21.1  Mean   :86.3  Mean   :17.5
3rd Qu.:62.0  3rd Qu.:24.0  3rd Qu.:89.0  3rd Qu.:19.0
Max.   :80.0  Max.   :27.0  Max.   :93.0  Max.   :42.0
```

We can compute these numbers separately also:

```
> stackloss$Air.Flow
[1] 80 80 75 62 62 62 62 62 58 58 58 58 58 58 50 50 50 50 50 56 70
> mean(stackloss$Ai)
[1] 60.429
> median(stackloss$Ai)
[1] 58
> range(stackloss$Ai)
[1] 50 80
> quantile(stackloss$Ai)
 0%  25%  50%  75% 100%
 50  56  58  62  80
```

We can get the variance and sd:

```
> var(stackloss$Ai)
[1] 84.057
> sqrt(var(stackloss$Ai))
[1] 9.1683
```

We can write a function to compute sd's:

```
> sd <- function(x) sqrt(var(x))
> sd(stackloss$Ai)
[1] 9.1683
```

We might also want the correlations:

```
> cor(stackloss)
           Air.Flow Water.Temp Acid.Conc. stack.loss
Air.Flow   1.00000   0.78185   0.50014   0.91966
Water.Temp 0.78185   1.00000   0.39094   0.87550
Acid.Conc. 0.50014   0.39094   1.00000   0.39983
stack.loss 0.91966   0.87550   0.39983   1.00000
```

Another numerical summary with a graphical element is the stem plot:

```
> stem(stackloss$Ai)
```

The decimal point is 1 digit(s) to the right of the |

```
5 | 000006888888
6 | 22222
7 | 05
8 | 00
```

## C.3 Graphical Summaries

We can make histograms and boxplot and specify the labels if we like:

```
> hist(stackloss$Ai)
> hist(stackloss$Ai,main="Histogram of Air Flow",
  xlab="Flow of cooling air")
> boxplot(stackloss$Ai)
```

Scatterplots are also easily constructed:

```
> plot(stackloss$Ai,stackloss$W)
> plot(Water.Temp ~ Air.Flow,stackloss,xlab="Air Flow",
  ylab="Water Temperature")
```

We can make a scatterplot matrix:

```
> plot(stackloss)
```

We can put several plots in one display

```
> par(mfrow=c(2,2))
> boxplot(stackloss$Ai)
> boxplot(stackloss$Wa)
> boxplot(stackloss$Ac)
> boxplot(stackloss$s)
> par(mfrow=c(1,1))
```

## C.4 Selecting subsets of the data

Second row:

```
> stackloss[2,]
  Air.Flow Water.Temp Acid.Conc. stack.loss
2         80         27         88         37
```

Third column:

```
> stackloss[,3]
 [1] 89 88 90 87 87 87 93 93 87 80 89 88 82 93 89 86 72 79 80 82 91
```

The 2,3 element:

```
> stackloss[2,3]
 [1] 88
```

`c()` is a function for making vectors, e.g.

```
> c(1,2,4)
 [1] 1 2 4
```

Select the first, second and fourth rows:

```
> stackloss[c(1,2,4),]
  Air.Flow Water.Temp Acid.Conc. stack.loss
1      80         27         89         42
2      80         27         88         37
4      62         24         87         28
```

The : operator is good for making sequences e.g.

```
> 3:11
[1] 3 4 5 6 7 8 9 10 11
```

We can select the third through sixth rows:

```
> stackloss[3:6,]
  Air.Flow Water.Temp Acid.Conc. stack.loss
3      75         25         90         37
4      62         24         87         28
5      62         22         87         18
6      62         23         87         18
```

We can use "-" to indicate "everything but", e.g. all the data except the first two columns is:

```
> stackloss[,-c(1,2)]
  Acid.Conc. stack.loss
1          89         42
2          88         37
... stuff deleted ...
21         91         15
```

We may also want select the subsets on the basis of some criterion e.g. which cases have an air flow greater than 72.

```
> stackloss[stackloss$Ai > 72,]
  Air.Flow Water.Temp Acid.Conc. stack.loss
1      80         27         89         42
2      80         27         88         37
3      75         25         90         37
```

## C.5 Learning more about R

While running R you can get help about a particular commands - eg - if you want help about the `stem()` command just type `help(stem)`.

If you don't know what the name of the command is that you want to use then type:

```
help.start()
```

and then browse. You may be able to learn the language simply by example in the text and referring to the help pages.

You can also buy the books mentioned in the recommendations or download various guides on the web — anything written for S-plus will also be useful.