Getting started with R

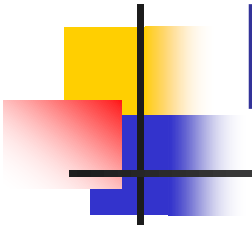
R Overview

R is a comprehensive statistical and graphical programming language and is a dialect of the S language:

1988 - S2: RA Becker, JM Chambers, A Wilks
1992 - S3: JM Chambers, TJ Hastie
1998 - S4: JM Chambers

R: initially written by Ross Ihaka and Robert Gentleman at Dep. of Statistics of U of Auckland, New Zealand during 1990s.

Since 1997: international "R-core" team of 15 people with access to common CVS archive.



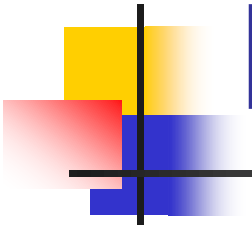
R Overview

You can enter commands one at a time at the command prompt (>) or run a set of commands from a source file.

There is a wide variety of data types, including vectors (numerical, character, logical), matrices, dataframes, and lists.

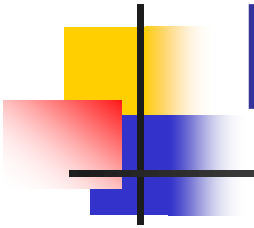
To quit R, use

>q()



R Overview

- Most functionality is provided through built-in and user-created functions and all data objects are kept in memory during an interactive session.
- Basic functions are available by default. Other functions are contained in packages that can be attached to a current session as needed



R Overview

A key skill to using **R** effectively is learning how to use the built-in help system. Other sections describe the working environment, inputting programs and outputting results, installing new functionality through packages and etc.

A fundamental design feature of **R** is that the output from most functions can be used as input to other functions. This is described in reusing results.

R Introduction

- Results of calculations can be stored in objects using the assignment operators:
  - An arrow (<-) formed by a smaller than character and a hyphen without a space!
  - The equal character (=).

R Introduction

- These objects can then be used in other calculations. To print the object just enter the name of the object. There are some restrictions when giving an object a name:
  - Object names cannot contain `strange' symbols like !, +, -, #.
  - A dot (.) and an underscore ( ) are allowed, also a name starting with a dot.
  - Object names can contain a number but cannot start with a number.
  - R is case sensitive, X and x are two different objects, as well as temp and temP.

An example

```
> # An example
> x <- c(1:10)
> x[(x>8) | (x<5)]
> # yields 1 2 3 4 9 10
> # How it works
> x <- c(1:10)
> X
>1 2 3 4 5 6 7 8 9 10
> x > 8
> F F F F F F F F T T
> x < 5
> T T T T F F F F F F
> x > 8 | x < 5
> T T T T F F F F T T
> x[c(T,T,T,T,F,F,F,F,T,T)]
> 1 2 3 4 9 10
```

R Introduction

- To list the objects that you have in your current R session use the function ls or the function objects.

```
> ls()
[1] "x" "y"
```

- So to run the function ls we need to enter the name followed by an opening (and and aclosing). Entering only ls will just print the object, you will see the underlying R code of the the function ls. Most functions in R accept certain arguments. For example, one of the arguments of the function ls is pattern. To list all objects starting with the letter x:

```
> x2 = 9
> y2 = 10
> ls(pattern="x")
[1] "x" "x2"
```

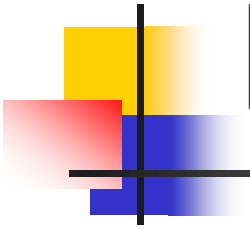
R Introduction

- If you assign a value to an object that already exists then the contents of the object will be overwritten with the new value (without a warning!). Use the function rm to remove one or more objects from your session.

```
> rm(x, x2)
```

- Lets create two small vectors with data and a scatterplot.

```
z2 <- c(1,2,3,4,5,6)
z3 <- c(6,8,3,5,7,1)
plot(z2,z3)
title("My first scatterplot")
```



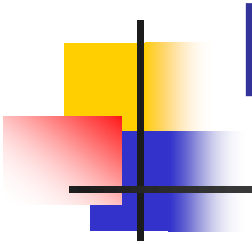
R Warning !

R is a case sensitive language.

FOO, Foo, and foo are three different objects

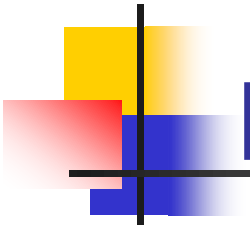
R Introduction

```
> x = sin(9)/75
> y = log(x) + x^2
> x
[1] 0.005494913
> y
[1] -5.203902
> m <- matrix(c(1,2,4,1), ncol=2)
> m
> [,1] [,2]
[1,] 1 4
[2,] 2 1
> solve(m)
[,1] [,2]
[1,] -0.1428571 0.5714286
[2,] 0.2857143 -0.1428571
```



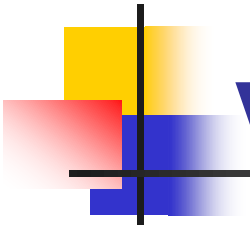
Data Input

- Data Types
- Importing Data
- Keyboard Input
- Database Input
- Exporting Data
- Viewing Data
- Variable Labels
- Value Labels
- Missing Data
- Date Values



Data Types

R has a wide variety of data types including scalars, vectors (numerical, character, logical), matrices, dataframes, and lists.

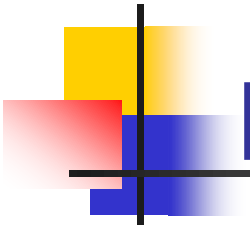


Vectors

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE)
#logical vector
```

Refer to elements of a vector using subscripts.

```
a[c(2,4)] # 2nd and 4th elements of vector
```



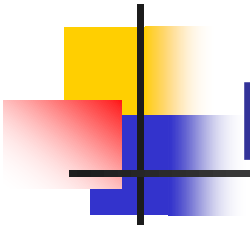
Matrices

All columns in a matrix must have the same mode(numeric, character, etc.) and the same length.

The general format is

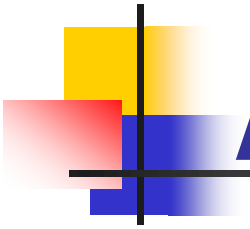
mymatrix <- **matrix(***vector*, **nrow=***r*, **ncol=***c*, **byrow=***FALSE*,**dimnames=list(***char_vector_rownames*, *char_vector_colnames***))**

byrow=TRUE indicates that the matrix should be filled by rows. **byrow=FALSE** indicates that the matrix should be filled by columns (the default). **dimnames** provides optional labels for the columns and rows.



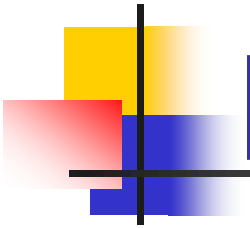
Matrices

```
# generates 5 x 4 numeric matrix
   y<-matrix(1:20, nrow=5,ncol=4)

# another example
   cells <- c(1,26,24,68)
   rnames <- c("R1", "R2")
   cnames <- c("C1", "C2")
   mymatrix <- matrix(cells, nrow=2, ncol=2,
   byrow=TRUE, dimnames=list(rnames, cnames))

#Identify rows, columns or elements using subscripts.

x[,4] # 4th column of matrix
   x[3,] # 3rd row of matrix
   x[2:4,1:3] # rows 2,3,4 of columns 1,2,3
```



Arrays

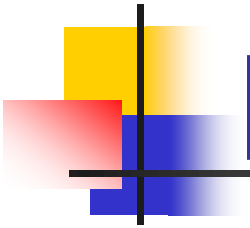
Arrays are similar to matrices but can have more than two dimensions. See **help(array)** for details.



Data frames

A data frame is more general than a matrix, in that different columns can have different modes (numeric, character, factor, etc.).

```
d <- c(1,2,3,4)
e <- c("red", "white", "red", NA)
f <- c(TRUE,TRUE,TRUE,FALSE)
mydata <- data.frame(d,e,f)
names(mydata) <- c("ID","Color","Passed")
   #variable names
```



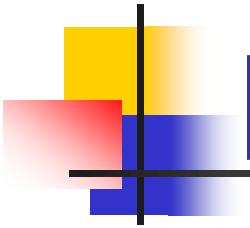
Data frames

There are a variety of ways to identify the elements of a dataframe .

myframe[3:5] # columns 3,4,5 of dataframe

myframe[c("ID","Age")] # columns ID and Age from dataframe

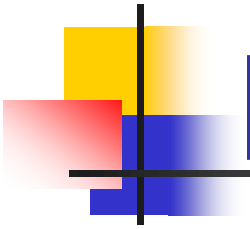
myframe$X1 # variable x1 in the dataframe



Lists

An ordered collection of objects (components). A list allows you to gather a variety of (possibly unrelated) objects under one name.

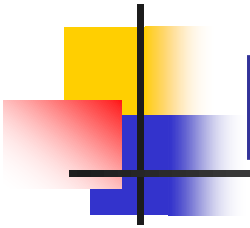
```
# example of a list with 4 components -
  # a string, a numeric vector, a matrix, and a scaler
  w <- list(name="Fred", mynumbers=a,
  mymatrix=y, age=5.3)

  # example of a list containing two lists
  v <- c(list1,list2)
```



Lists

Identify elements of a list using the [[]] convention.

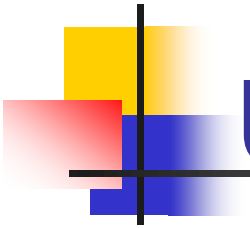
```
mylist[[2]] # 2nd component of the list
```



Factors

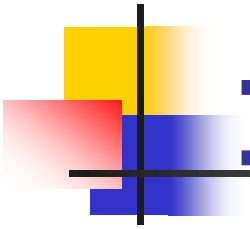
Tell **R** that a variable is **nominal** by making it a factor. The factor stores the nominal values as a vector of integers in the range [1... k] (where k is the number of unique values in the nominal variable), and an internal vector of character strings (the original values) mapped to these integers.

```
# variable gender with 20 "male" entries and
# 30 "female" entries
  gender <- c(rep("male",20), rep("female", 30))
  gender <- factor(gender)
# stores gender as 20 1s and 30 2s and associates
# 1=female, 2=male internally (alphabetically)
# R now treats gender as a nominal variable
  summary(gender)
```



Useful Functions

```
length(object) # number of elements or components
str(object)    # structure of an object
class(object)  # class or type of an object
names(object)  # names
c(object,object,...)  # combine objects into a vector
cbind(object, object, ...) # combine objects as columns
rbind(object, object, ...) # combine objects as rows
ls()       # list current objects
rm(object) # delete an object
newobject <- edit(object) # edit copy and save a
newobject
fix(object)               # edit in place
```



Importing Data

Importing data into **R** is fairly simple.

For Stata and Systat, use the **foreign** package.

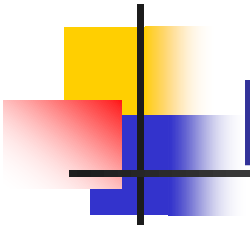
For SPSS and SAS I would recommend the **Hmisc** package for ease and functionality.

See the **Quick-R** section on **packages**, for information on obtaining and installing the these packages.

Example of importing data are provided below.

From A Comma Delimited Text File

```
# first row contains variable names, comma is
  separator
# assign the variable id to row names
# note the / instead of \ on mswindows systems

  mydata <- read.table("c:/mydata.csv",
  header=TRUE, sep=",", row.names="id")
```



From Excel

The best way to read an Excel file is to export it to a comma delimited file and import it using the method above.

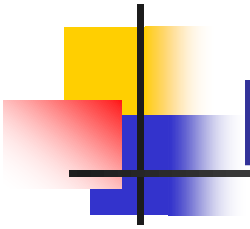
On windows systems you can use the **RODBC** package to access Excel files. The first row should contain variable/column names.

```
# first row contains variable names
# we will read in workSheet mysheet
library(RODBC)
channel <- odbcConnectExcel("c:/myexel.xls")
mydata <- sqlFetch(channel, "mysheet")
odbcClose(channel)
```

From SAS

- # save SAS dataset in trasport format
  libname out xport 'c:/mydata.xpt';
  data out.mydata;
  set sasuser.mydata;
  run;
- library(foreign)

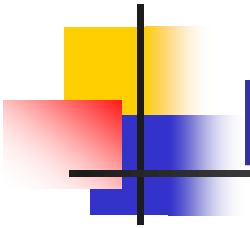
  #bsl=read.xport("mydata.xpt")



Keyboard Input

Usually you will obtain a dataframe by importing it from **SAS**, **SPSS**, **Excel**, **Stata**, a database, or an ASCII file. To create it interactively, you can do something like the following.

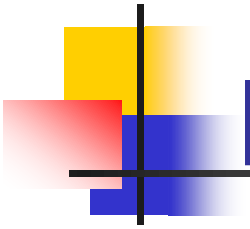
```
# create a dataframe from scratch
age <- c(25, 30, 56)
gender <- c("male", "female", "male")
weight <- c(160, 110, 220)
mydata <- data.frame(age,gender,weight)
```



Keyboard Input

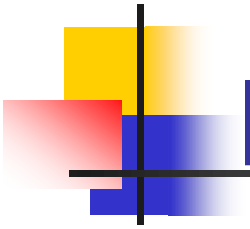
You can also use **R**'s built in spreadsheet to enter the data interactively, as in the following example.

```
# enter data using editor
mydata <- data.frame(age=numeric(0),
gender=character(0), weight=numeric(0))
mydata <- edit(mydata)
# note that without the assignment in the line
above,
# the edits are not saved!
```



Exporting Data

There are numerous methods for exporting **R** objects into other formats . For SPSS, SAS and Stata. you will need to load the **foreign** packages. For Excel, you will need the **xlsReadWrite** package.



Exporting Data

To A Tab Delimited Text File

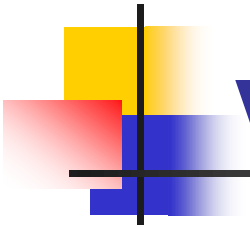
```
write.table(mydata, "c:/mydata.txt", sep="\t")
```

To an Excel Spreadsheet

```
library(xlsReadWrite)
   write.xls(mydata, "c:/mydata.xls")
```

To SAS

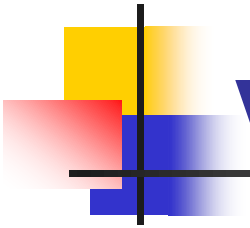
```
library(foreign)
   write.foreign(mydata, "c:/mydata.txt",
   "c:/mydata.sas",   package="SAS")
```



Viewing Data

There are a number of functions for listing the contents of an object or dataset.

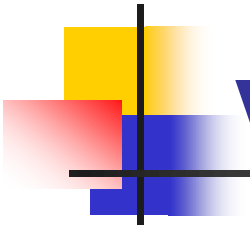
```
# list objects in the working environment
ls()
# list the variables in mydata
names(mydata)
# list the structure of mydata
str(mydata)
# list levels of factor v1 in mydata
levels(mydata$v1)
# dimensions of an object
dim(object)
```



Viewing Data

There are a number of functions for listing the contents of an object or dataset.

```
# class of an object (numeric, matrix, dataframe, etc)
  class(object)
# print mydata
  mydata
# print first 10 rows of mydata
  head(mydata, n=10)
# print last 5 rows of mydata
  tail(mydata, n=5)
```

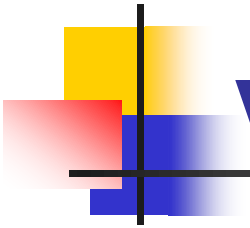


Variable Labels

R's ability to handle variable labels is somewhat unsatisfying.

If you use the **Hmisc** package, you can take advantage of some labeling features.

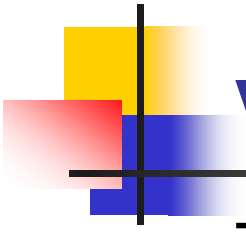
```
library(Hmisc)
label(mydata$myvar) <- "Variable label for variable myvar"
describe(mydata)
```



Variable Labels

Unfortunately the label is only in effect for functions provided by the **Hmisc** package, such as **describe()**. Your other option is to use the variable label as the variable name and then refer to the variable by position index.

```
names(mydata)[3] <- "This is the label for variable 3"
mydata[3] # list the variable
```

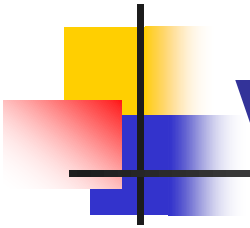


Value Labels

To understand value labels in **R**, you need to understand the data structure factor.

You can use the factor function to create your own value lables.

```
# variable v1 is coded 1, 2 or 3
# we want to attach value labels 1=red, 2=blue,3=green
  mydata$v1 <- factor(mydata$v1,
  levels = c(1,2,3),
  labels = c("red", "blue", "green"))
# variable y is coded 1, 3 or 5
# we want to attach value labels 1=Low, 3=Medium, 5=High
```

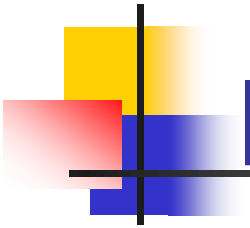


Value Labels

```
mydata$v1 <- ordered(mydata$y,
    levels = c(1,3, 5),
    labels = c("Low", "Medium", "High"))
```

Use the **factor()** function for **nominal data** and the **ordered()** function for **ordinal data**. **R** statistical and graphic functions will then treat the data appropriately.

Note: factor and ordered are used the same way, with the same arguments. The former creates factors and the later creates ordered factors.

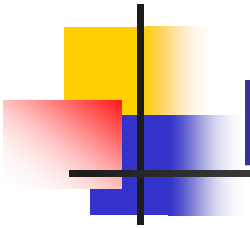


Missing Data

In **R**, missing values are represented by the symbol **NA** (not available) . Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number). Unlike SAS, **R** uses the same symbol for character and numeric data.

Testing for Missing Values

```
is.na(x) # returns TRUE of x is missing
y <- c(1,2,3,NA)
is.na(y) # returns a vector (F F F T)
```



Missing Data

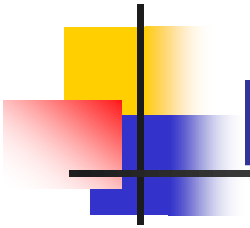
Recoding Values to Missing

```
# recode 99 to missing for variable v1
# select rows where v1 is 99 and recode column v1
mydata[mydata$v1==99,"v1"] <- NA
```

Excluding Missing Values from Analyses

Arithmetic functions on missing values yield missing values.

```
x <- c(1,2,NA,3)
mean(x)           # returns NA
mean(x, na.rm=TRUE) # returns 2
```



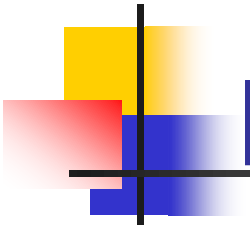
Missing Data

The function **complete.cases()** returns a logical vector indicating which cases are complete.

```
# list rows of data that have missing values
mydata[!complete.cases(mydata),]
```

The function **na.omit()** returns the object with listwise deletion of missing values.

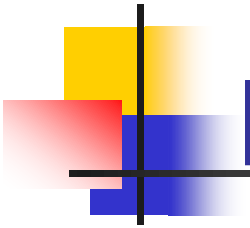
```
# create new dataset without missing data
newdata <- na.omit(mydata)
```



Missing Data

Advanced Handling of Missing Data

Most modeling functions in **R** offer options for dealing with missing values. You can go beyond pairwise of listwise deletion of missing values through methods such as multiple imputation. Good implementations that can be accessed through **R** include **Amelia II**, **Mice**, and **mitools**.



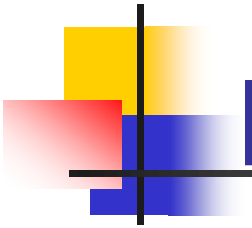
Date Values

Dates are represented as the number of days since 1970-01-01, with negative values for earlier dates.

```
# use as.Date( ) to convert strings to dates
mydates <- as.Date(c("2007-06-22", "2004-02-13"))
# number of days between 6/22/07 and 2/13/04
    days <- mydates[1] - mydates[2]
```

Sys.Date( ) returns today's date.

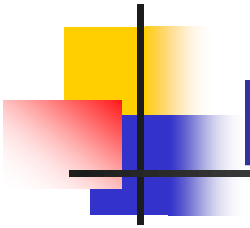
Date() returns the current date and time.



Date Values

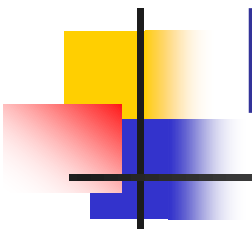
The following symbols can be used with the format( ) function to print dates.

Symbol	Meaning	Example
%d	day as a number (0-31)	01-31
%a	abbreviated weekday	Mon
%A	unabbreviated weekday	Monday
%m	month (00-12)	00-12
%b	abbreviated month	Jan
%B	unabbreviated month	January
%y	2-digit year	07
%Y	4-digit year	2007



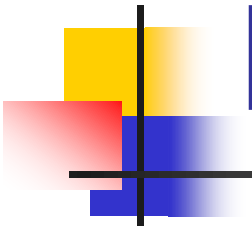
Date Values

```
# print today's date
today <- Sys.Date()
format(today, format="%B %d %Y")
     "June 20 2007"
```



R Workspace

Objects that you create during an R session are hold in memory, the collection of objects that you currently have is called the workspace. This workspace is not saved on disk unless you tell R to do so. This means that your objects are lost when you close R and not save the objects, or worse when R or your system crashes on you during a session.



R Workspace

When you close the RGui or the R console window, the system will ask if you want to save the workspace image. If you select to save the workspace image then all the objects in your current R session are saved in a file .RData. This is a binary file located in the working directory of R, which is by default the installation directory of R.

R Workspace

- During your R session you can also explicitly save the workspace image. Go to the `File' menu and then select `Save Workspace...', or use the save.image function.

```
## save to the current working directory
save.image()
## just checking what the current working directory is
getwd()
## save to a specific file and location
save.image("C:\\Program Files\\R\\R-
   2.5.0\\bin\\.RData")
```

R Workspace

If you have saved a workspace image and you start R the next time, it will restore the workspace. So all your previously saved objects are available again. You can also explicitly load a saved workspace le, that could be the workspace image of someone else. Go the `File' menu and select `Load workspace...'.

R Workspace

Commands are entered interactively at the **R** user prompt. **Up** and **down arrow keys** scroll through your command history.

You will probably want to keep different projects in different physical directories.

R Workspace

R gets confused if you use a path in your code like

c:|mydocuments|myfile.txt

This is because R sees "\" as an escape character. Instead, use

c:||my documents||myfile.txt

or

c:/mydocuments/myfile.txt

R Workspace

```
getwd() # print the current working directory

ls()  # list the objects in the current workspace
setwd(mydirectory)   # change to mydirectory
setwd("c:/docs/mydir")
```

R Workspace

```
#view and set options for the session
  help(options) # learn about available options
  options() # view current option settings
  options(digits=3) # number of digits to print
  on output
# work with your previous commands
  history() # display last 25 commands
  history(max.show=Inf) # display all previous commands
```

R Workspace

```
# save your command history
  savehistory(file="myfile") # default is ".Rhistory"
# recall your command history
  loadhistory(file="myfile") # default is ".Rhistory"
```

R Help

Once **R** is installed, there is a comprehensive built-in help system. At the program's command prompt you can use any of the following:

```
help.start()    # general help
help(foo)       # help about function foo
?foo            # same thing
apropos("foo") # list all function containing string foo
example(foo)    # show an example of function foo
```

R Help

```
# search for foo in help manuals and archived mailing lists
   RSiteSearch("foo")
# get vignettes on using installed packages
   vignette()      # show available vingettes
   vignette("foo") # show specific vignette
```

R Datasets

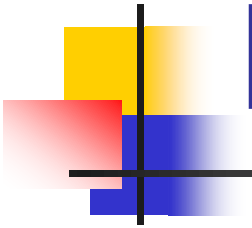
R comes with a number of sample datasets that you can experiment with. Type

> data( )

to see the available datasets. The results will depend on which packages you have loaded. Type

help(*datasetname***)**

for details on a sample dataset.



R Packages

- One of the strengths of R is that the system can easily be extended. The system allows you to write new functions and package those functions in a so called `R package' (or `R library'). The R package may also contain other R objects, for example data sets or documentation. There is a lively R user community and many R packages have been written and made available on CRAN for other users. Just a few examples, there are packages for portfolio optimization, drawing maps, exporting objects to html, time series analysis, spatial statistics and the list goes on and on.

R Packages

- When you download R, already a number (around 30) of packages are downloaded as well. To use a function in an R package, that package has to be attached to the system. When you start R not all of the downloaded packages are attached, only seven packages are attached to the system by default. You can use the function search to see a list of packages that are currently attached to the system, this list is also called the search path.

```
> search()
[1] ".GlobalEnv" "package:stats" "package:graphics"
[4] "package:grDevices" "package:datasets" "package:utils"
[7] "package:methods" "Autoloads" "package:base"
```

R Packages

To attach another package to the system you can use the menu or the library function. Via the menu:

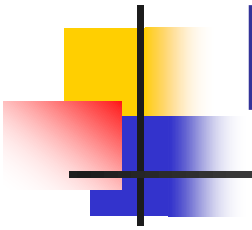
Select the `Packages' menu and select `Load package...', a list of available packages on your system will be displayed. Select one and click `OK', the package is now attached to your current R session. Via the library function:

```
> library(MASS)
> shoes
$A
[1] 13.2 8.2 10.9 14.3 10.7 6.6 9.5 10.8 8.8 13.3
$B
[1] 14.0 8.8 11.2 14.2 11.8 6.4 9.8 11.3 9.3 13.6
```

R Packages

- The function library can also be used to list all the available libraries on your system with a short description. Run the function without any arguments

```
> library()
Packages in library C:/Program Files/R/R-3.2.3/library':
base                    The R Base Package
Boot                     Bootstrap R (S-Plus) Functions  (Canty)
class                   Functions for Classification
cluster                 Cluster Analysis Extended Rousseeuw  et al.
codetools               Code Analysis Tools for R
datasets                The R Datasets Package
DBI                     R Database Interface
foreign                 Read Data Stored by Minitab, S, SAS,
                SPSS, Stata, Systat, dBase, ...
graphics                The R Graphics Package
```



R Packages

```
install = function() {
install.packages(c("moments","graphics","Rcmdr","hexbin"),
repos="http://lib.stat.cmu.edu/R/CRAN")
}
install()
```

R Conflicting objects

- It is not recommended to do, but R allows the user to give an object a name that already exists. If you are not sure if a name already exists, just enter the name in the R console and see if R can find it. R will look for the object in all the libraries (packages) that are currently attached to the R system. R will not warn you when you use an existing name.

```
> mean = 10
> mean
[1] 10
```

- The object mean already exists in the base package, but is now masked by your object mean. To get a list of all masked objects use the function conflicts.

```
>
[1] "body<-" "mean"
```

R Conflicting objects

The object mean already exists in the base package, but is now masked by your object mean. To get a list of all masked objects use the function conflicts.

```
> conflicts()
[1] "body<-" "mean"
```

You can safely remove the object mean with the function rm() without risking deletion of the mean function.

Calling rm() removes only objects in your working environment by default.

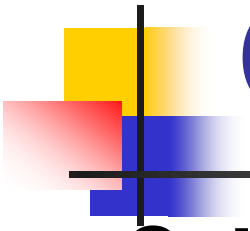
Source Codes

you can have input come from a script file (a file containing **R** commands) and direct output to a variety of destinations.

Input

The **source( )** function runs a script in the current session. If the filename does not include a path, the file is taken from the current working directory.

```
# input a script
source("myfile")
```

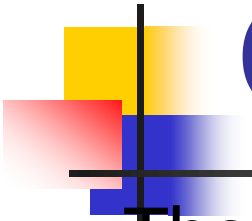


Output

Output

The **sink( )** function defines the direction of the output.

```
# direct output to a file
sink("myfile", append=FALSE, split=FALSE)

# return output to the terminal
sink()
```



Output

The **append** option controls whether output overwrites or adds to a file.

The **split** option determines if output is also sent to the screen as well as the output file.

Here are some examples of the **sink()** function.

```
# output directed to output.txt in c:\projects directory.
# output overwrites existing file. no output to terminal.
sink("myfile.txt", append=TRUE, split=TRUE)
```

Graphs

To redirect graphic output use one of the following functions. Use **dev.off( )** to return output to the terminal.

Function	Output to
pdf("mygraph.pdf")	pdf file
win.metafile("mygraph.wmf")	windows metafile
png("mygraph.png")	png file
jpeg("mygraph.jpg")	jpeg file
bmp("mygraph.bmp")	bmp file
postscript("mygraph.ps")	postscript file

Redirecting Graphs

```
# example - output graph to jpeg file
  jpeg("c:/mygraphs/myplot.jpg")
  plot(x)
  dev.off()
```

Reusing Results

One of the most useful design features of **R** is that the output of analyses can easily be saved and used as input to additional analyses.

```
# Example 1
lm(mpg~wt, data=mtcars)
```

This will run a simple linear regression of miles per gallon on car weight using the dataframe mtcars. Results are sent to the screen. Nothing is saved.

Reusing Results

```
# Example 2
  fit <- lm(mpg~wt, data=mtcars)
```

This time, the same regression is performed but the results are saved under the name fit. No output is sent to the screen. However, you now can manipulate the results.

```
str(fit) # view the contents/structure of "fit"
```

The assignment has actually created a list called "fit" that contains a wide range of information (including the predicted values, residuals, coefficients, and more.

Reusing Results

```
# plot residuals by fitted values
  plot(fit$residuals, fit$fitted.values)
```

To see what a function returns, look at the **value** section of the online help for that function. Here we would look at **help(lm)**.

The results can also be used by a wide range of other functions.

```
# produce diagnostic plots
  plot(fit)
```