Week 1 Lab - Introduction to R

Unit Coordinator - Dr Liwan Liyanage

School of Computing, Engineering and Mathematics

WESTERN SYDNEY UNIVERSITY

What is R?

R is a software environment for statistical computing and graphics. It runs on just about any platform (except iPad!) and is completely free (in the GNU sense).

It is used extensively by academic statisticians for research and teaching and is gaining ground in business.

It has 4634 extension packages available.

Pros

Its free and open source. It has most methods for most things mostly before any other package. It has the best graphics. It extendable.

Cons

It has a steep learning curve. No GUI by default. Poor (but improving) memory management; difficulty with very large data sets.

R Resources

- http://www.r-project.org - Main R website.
- CRAN - http://cran.csiro.au - Comprehensive R Archive Network - base software and add-on packages.
- RStudio - http://www.rstudio.com - is a powerful IDE for R
- R Commander - install.package(Rcmdr) - is a partial GUI interface to R - requires TclTk.
- R Graph Gallery - http://gallery.r-enthusiasts.com/ - loads of pretty pictures.
- http://cran.csiro.au/doc/contrib/Torfs+ Brauer-Short-R-Intro.pdf - “A (very) short Introduction to R”
- “Introductory Statistics with R”, Peter Dalgaard, Springer 2008.

R Commands

R can be used as a basic calculator.

```
1+1
```

```
## [1] 2
```

```
sqrt(2)
```

```
## [1] 1.414214
```

```
2^5
```

```
## [1] 32
```

R Commands ctd. . .

It can store and print variables.

```r
x=1
print(x)
```

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

WESTERN SYDNEY
UNIVERSITY

R Commands ctd…

It understands vectors and matrices.

```
x <- c(1,2)
m <- matrix(c(1,2,3,4), ncol=2, byrow=TRUE)
print(m)
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

R Commands ctd...

It has functions, and you can write them.

```r
x <- sqrt(2)
sqr <- function(x) x^2
sqr(2)
```

```
## [1] 4
```

Uploading Data Into R

iris<- read.csv("C:/LIWAN/R/2016/Intro to Data Science/iris.csv")
attach(iris)

Data can be read from text files (read.csv and read.table) and various formats using the foreign package.For example; dataset = read.csv("MyData.csv")

When the data set is uploaded to the same same folder where R project is saved, use

iris<- read.csv("iris.csv") attach(iris)

Data in R

Tables are stored in data.frames

```r
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.2  setosa
```

```r
sapply(iris,class)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width    Sp
##    "numeric"    "numeric"    "numeric"    "numeric"    "fa
```

Summary

```
names(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Wi
## [5] "Species"
```

```
dim(iris)
```

```
## [1] 150   5
```

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Wi
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.
##  Median :5.800   Median :3.000   Median :4.350   Median :1.
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.
```

Basic Statistics

```r
x<-rnorm(100)
mean(x)
```

```
## [1] -0.008156444
```

```r
var(x)
```

```
## [1] 1.284828
```

```r
sd(x)
```

```
## [1] 1.133503
```

```r
fivenum(x)
```

```
## [1] -3.09103456 -0.71836398 -0.02218772  0.67031270  3.0023
```

minimum lower quartile median upper quartile maximum

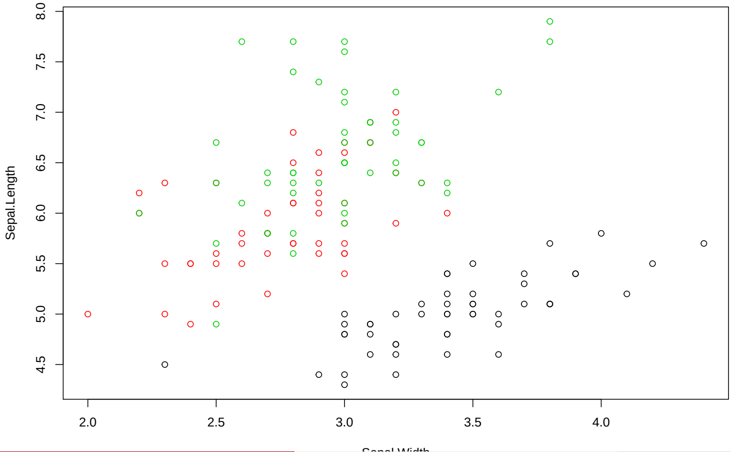
Basic Statistics

```r
t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -0.071958, df = 99, p-value = 0.9428
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2330680  0.2167551
## sample estimates:
##    mean of x
## -0.008156444
```

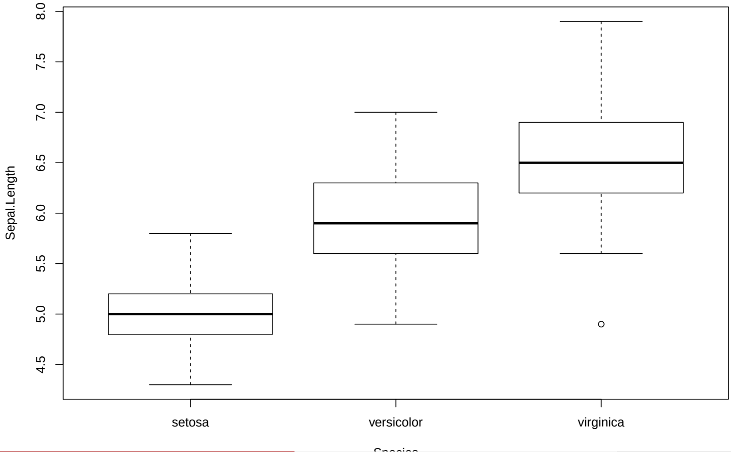
R has extensive plotting

```
plot(Sepal.Length~Sepal.Width, col=Species, data=iris)
```



R has extensive plotting

```r
boxplot(Sepal.Length~Species, data=iris)
```



Help in R

Everything in R has a help file.

help(t.test)

Or see the help pane in RStudio

Will illustrate R further within Regression Analysis

Getting Ready for the Data Analysis covered in Lectures

Data Import

```
install.packages(“ISLR”)
install.packages(“MASS”)
library(ISLR)
library(MASS)
library(class)
library(DMwR)
attach(Smarket)
attach(Boston)
attach(Carseats)
attach(iris)
```

Data Sets

Supervised Learning:

- Advertising
- Income
- Heart
- Smarket
- Caravan (Insurance Data)

Unsupervised Learning:

- USAarrests
- groceries

View Advertising Data and Discuss How to Initiate Knowledge Discovery

Exercise: List Possible Researcch Questions?

```
Advertising<-read.csv("Advertising.csv")
attach(Advertising)
head(Advertising)
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

View heart Data and Discuss How to Initiate Knowledge Discovery

Exercise: List Possible Researcch Questions?

```r
Heart<-read.csv("heart.csv")
attach(Heart)
head(Heart)
```

```
##   X Age Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAn
## 1 1  63   1      typical    145  233   1       2   150
## 2 2  67   1 asymptomatic    160  286   0       2   108
## 3 3  67   1 asymptomatic    120  229   0       2   129
## 4 4  37   1   nonanginal    130  250   0       0   187
## 5 5  41   0   nontypical    130  204   0       2   172
## 6 6  56   1   nontypical    120  236   0       0   178
##   Ca       Thal AHD
## 1  0      fixed   0
## 2  3     normal   1
```

View groceries Data and Discuss How to Initiate Knowledge Discovery

Exercise: List Possible Researcch Questions?

```r
Groceries<-read.csv("groceries.csv")
attach(Groceries)
head(Groceries)
```

```
##   frankfurter sausage liver.loaf ham meat finished.product
## 1           0       0          0   0    0                0
## 2           0       0          0   0    0                0
## 3           0       0          0   0    0                0
## 4           0       0          0   0    0                0
## 5           0       0          0   0    0                0
## 6           0       0          0   0    0                0
##   organic.sausage chicken turkey pork beef hamburger.meat f
## 1               0       0      0    0    0              0
## 2               0       0      0    0    0              0
```

Explore Default Data set from the ISLR Library

```r
#install.packages("ISLR")
library(ISLR)
attach(Default)
View(Default)
dim(Default)
```

```
## [1] 10000     4
```

```r
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

Save and read datasets within R

Save a data set downloaded from a library within R as a csv file

```
write.csv(Default,file="Default.csv")
```

To read csv files

read.csv(“Default.csv”,header=TRUE)

To read any other files read.table(“file”,header=False)

How to change a factor variable to a numeric variable

Add another variable named Defcode to table Default check the levels of the new variable (It will be same class as the original variable)

```
Defcode = Default$default
levels(Defcode)
```

```
## [1] "No"  "Yes"
```

Change the levels as 1 for Yes and 0 for No

```
levels(Defcode)[levels(Defcode)=="No"]=0
levels(Defcode)[levels(Defcode)=="Yes"]=1
levels(Defcode)
```

```
## [1] "0" "1"
```

Still Defcode varibale is a factor variable and cannot use as a numeric variable in regression setting.

To summariase a factor variable

```r
Defcode = as.character(Default$default)
table(Defcode)
```

```
## Defcode
##   No  Yes
## 9667  333
```

```r
Defcode[Defcode=="No"]=0
Defcode[Defcode=="Yes"]=1
table(Defcode)
```

```
## Defcode
##    0    1
## 9667  333
```

Change a factor vraiable to a numeric variable

```
Defcode = as.numeric(Defcode)
class(Defcode)
```

```
## [1] "numeric"
```

Exercises

For each of the three data sets, iris, heart and groceries

- Explore the variables
- List the quantitative variables and qualitative variables
- State a Research question and identify the target variable if aplicable
- Comment if they are supervised learning or unsupervised learning.