1 Class Assignment: Introduction to R

1.1 Introduction

Using this sample code,

```
install.packages("BB")

library(BB)

source("A1.R")

?for

??rpareto

dir()

1+1

2/2

save.image("misc.RDATA")

1:10

30%%4

setwd("/Users/ms486/Dropbox/Papers/Progress")

getwd()

ls()

2/0

log(-1)

sum(1:10)
```

Exercise 1 Introduction

1. Create a directory for this class and store your script “a0.R”

2. Install the packages, Hmisc, gdata,boot,xtable,MASS,moments,snow,mvtnorm

3. Set your working directory

4. List the content of your directory and the content of your environment

5. Check whether 678 is a multiple of 9

6. Save your environment
7. Find help on the function mean, cut2
8. Find an operation that returns NaN (Not A Number)

1.2 Objects

Vectors, Matrix, Arrays

```
vec0  = NULL
vec1  = c(1,2,3,4)
vec2  = 1:4
vec3  = seq(1,4,1)
vec4  = rep(0,4)
sum(vec1)
str(vec1)
prod(vec1)
mat1  = mat.or.vec(2,2)
mat2  = matrix(0,ncol=2,nrow=2,byrow=T)
mat3  = cbind(c(0,0),c(0,0))
mat4  = rbind(c(1,1),c(0,0))
mat5  = matrix(1:20,nrow=5,ncol=4)
mat5[1:2,3:4]
mat5[1,]
arr1  = array(0,c(2,2))
dim(mat4)
dim(vec2)
length(vec2)
length(mat1)
class(mat4)
```

Exercise 2 Object Manipulation

1. Print Titanic, and write the code to answer these questions (one function (sum) , one operation)
    (a) Total population
    (b) Total adults
    (c) Total crew
    (d) 3^{rd} class children
    (e) 2^{nd} class adult female
    (f) 1^{st} class children male
    (g) Female Crew survivor
    (h) 1^{st} class adult male survivor
2. Using the function *prop.table*, find
    (a) The proportion of survivors among first class, male, adult
    (b) The proportion of survivors among first class, female, adult
    (c) The proportion of survivors among first class, male, children
    (d) The proportion of survivors among third class, female, adult

Exercise 3 Vectors - Introduction

1. Use three different ways, to create the vectors
    (a) $a = 1, 2, \ldots, 50$
    (b) $b = 50, 49, \ldots, 1$

   Hint : rev
2. Create the vectors
    (a) $a = 10, 19, 7, 10, 19, 7, \ldots, 10, 19, 7$ with 15 occurrences of 10,19,7
    (b) $b = 1, 2, 5, 6, \ldots, 1, 2, 5, 6$ with 8 occurrences of 1,2,5,6

   Hint : rep

3. Create a vector of the values of $log(x)sin(x)$ at $x = 3.1, 3.2, \ldots, 6$

4. Using the function sample, draw 90 values between (0,100) and calculate the mean. Re-do the same operation allowing for replacement.

5. Calculate

   (a) $\sum_{a=1}^{20} \sum_{b=1}^{15} \frac{exp(\sqrt{a})log(a^5)}{5 + cos(a)sin(b)}$

   (b) $\sum_{a=1}^{20} \sum_{b=1}^{a} \frac{exp(\sqrt{a})log(a^5)}{5 + exp(ab)cos(a)sin(b)}$

6. Create a vector of the values of $\exp(x)\cos(x)$ at x = 3, 3.1, ...6.

Exercise 4 Vectors - Advanced

1. Create two vectors $xVec$ and $yVec$ by sampling 1000 values between 0 and 999.

2. Suppose $xVec = (x_1, \ldots, x_n)$ and $yVec = (y_1, \ldots, y_n)$

   (a) Create the vector $(y_2 - x_1, \ldots, y_n - x_{n-1})$ denoted by $zVec.$

   (b) Create the vector $(\frac{sin(y_1)}{\cos(x_2}, \frac{sin(y_2)}{\cos(x_3}, \ldots, \frac{sin(y_{n-1})}{\cos(x_n})$ denoted by $zVec.$

   (c) Create a vector $subX$ which consists of the values of X which are $\geq$ 200.

   (d) What are the index positions in yVec of the values which are $\geq$ 600.

Exercise 5 Matrix

1. Create the matrix $A = \begin{vmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ -2 & -1 & -3 \end{vmatrix}$

   (a) Check that A^3=0 (matrix 0).

   (b) Bind a fourth column as the sum of the first and third column

   (c) Replace the third row by the sum of the first and second row

   (d) Calculate the average by row and column.

2. Consider this system of linear equations:

$$2x + y + 3z = 10 \tag{1}$$

$$x + y + z = 6 \tag{2}$$

$$x + 3y + 2z = 13 \tag{3}$$

3. Solve this equation.

Exercise 6 Functions

1. Write a function *fun1* which takes two arguments (a,n) where (a) is a scalar and n is a positive integer, and returns

$$a + \frac{a^2}{2} + \frac{a^3}{3} + \ldots + \frac{a^n}{n}$$

2. Consider the function

$$f(x) = \begin{cases} x^2 + 2x + |x| & \text{if } x < 0; \\ x^2 + 3 + \log(1 + x) & \text{if } 0 \leq x < 2; \\ x^2 + 4x - 14 & \text{if } x \geq 2. \end{cases} \tag{4}$$

Evaluate the function at -3, 0 and 3.

Exercise 7 Indexes

1. Sample 36 values between 1 and 20 and name it $v1$

2. Use two different ways, to create the subvector of elements that are not in the first row. *Hint: which and subset can not be used.* Check x[a] and x[-a].

3. Create a logical element (TRUE or FALSE), v2, which is true if $v1 > 5$. Can you convert this logical element into a dummy 1 (TRUE) and 0 (FALSE)?

4. Create a matrix m1 [6 × 6] which is filled by row using the vector v1.

5. Create the following object

```
x = c(rnorm(10),NA,paste("d",1:16),NA,log(rnorm(10)))
```

6. Test for the position of missing values, and non-finite values. Return a subvector free of missing and non-finite values.

Exercise 8 Data Manipulation

1. Load the library AER, and the dataset (data(”GSOEP9402”)) to be named dat.
2. What type of object is it? Find the number of rows and column? Can you provide the names of the variables?
3. Evaluate and plot the average annual income by year.
4. Create an array that illustrates simultaneously the income differences (mean) by gender, school and memployment.

Exercise 9 First regression

1. Load the dataset (data(”CASchools”)) to be named data.
2. Using the function lm, run a regression of read on the following variables: district, school, county, grades, students, teachers, calworks, lunch, computer, expenditure, income and english. Store this regression as reg1.
3. Can you run a similar regression by specifying,

   ```
   formula = y ~ x. lm(formula)
   ```

   Create reg2, that uses only the 200 first observations.

Exercise 10 Advanced indexing

1. Create a vector *lu* of 200 draws from a pareto distribution (1,1). How many values are higher than 10. Replace these values by draws from a logistic distribution (6.5,0.5).

2. Create a vector *de* of 200 draws from a normal distribution (1,2). Set $de = \log(de)$, and count the number of missing values or negative values. Replace these values by draws from a normal distribution (0,1) truncated at 0. *hint:truncnorm*

3. Create two vectors, *orig* and *dest* as 200 draws from a uniform distribution [0,1].

4. Create two matrices, *hist* and *dist* as 200*200 draws from a uniform distribution [0,1].

5. Consider this function

$$q_{jl}(w) = \frac{r+de_j}{r+de_l} w + lu_j log(w) - lu_l(1+log(w)) + \frac{r+de_j}{r+de_l}\sum_{k\neq j} su_{jk} - \sum_{k\neq l} su_{lk} + \frac{r+de_j}{r+de_l}\sum_{k\neq j} se_{jk} - \sum_{k\neq l} se_{lk} \tag{5}$$

   where

$$su_{j,l} = \log(orig_j + dest_l + dist_{j,l})/(1 + \log(orig_j + dest_l + dist_{j,l})) \tag{6}$$

$$se_{j,l} = \exp(orig_j + dest_l + hist_{j,l})/(1 + \exp(orig_j + dest_l + hist_{j,l})) \tag{7}$$

6. Create the matrices su and se.

7. Set r = 0.05. Create a function to evaluate $q_{jl}(.)$. Evaluate $q_{jl}(9245)$ for all pairs (j,l).

8. Create gridw, which consists of a sequence from 9100 to 55240 of length 50.

9. Using the function sapply, evaluate q_{jl}. Store the ouput into an array of dimension ($50 \times 200 \times 200$). How long does it take to evaluate $q_{jl}()$ for each value of w?

List

```
li      = list()

li[[1]] = mat1

li[[2]] = Titanic

li1     = list(x=mat1,y=Titanic)

li1$x
```

```
li2$y
```

Dataframe

```
data=data.frame(x=rnorm(100),y=runif(100))

data

browse(data)

edit(data)

data[,1]

data[1,]

data$x

names(data)

attach(data)

x

detach(data)

y
```

Tests and Conversion

```
is.na()

is.list()     as.list()

is.factor()   as.factor()

is.matrix()

is.vector()

is.array()

is.finite()

a==b

a=>b

a<=b
```

Exercise 11 Tests and indexing

1. Test if $c(1,2,3)$ is an array? a vector? a matrix?

2. x0 = rnorm(1000); Using the function *table()* count the number of occurrences of $x0 > 0$, $x0 > 1$, $x0 > 2$, $x0 > 0.5$, $x0 < 1$ and $x0 > -1$

3. `x1 = cut2(runif(100,0,1),g=10)`

   `levels(x1)=paste("q",1:10,sep="")`

4. Test whether or not x1 is a factor?

5. Verify that "q1" has 10 occurences.

6. Convert x1 into a numeric variables. What happens to the levels?

7. `rand = rnorm(1000)`

8. Using the function *which()* find the indexes of positive values.

9. Create the object w of positive values of x using:

   (a) Which

   (b) Subset

   (c) By indexing directly the values that respect a condition

1.3 Basic functions

Table 1: Basic Functions

Function	Description
abs(x)	absolute value
sqrt(x)	square root
ceiling(x)	ceiling(3.475) is 4
floor(x)	floor(3.475) is 3
trunc(x)	trunc(5.99) is 5
round(x, digits=n)	round(3.475, digits=2) is 3.48
signif(x, digits=n)	signif(3.475, digits=2) is 3.5
log(x)	logarithm
exp(x)	e^x
substr(x, start=n1, stop=n2)	Extract or replace substrings in a character vector. x = ”abcdef” , substr(x, 2, 4) is ”bcd”
grep(pattern, x)	Search for pattern in x.
sub(pattern, replacement, x)	Find pattern in x and replace with replacement text.
strsplit(x, split)	Split the elements of character vector x at split.
strsplit(”abc”, ””)	returns 3 element vector ”a”,”b”,”c”
paste(..., sep=””)	Concatenate strings
toupper(x)	Uppercase
tolower(x)	Lowercase

1.4 Language

```
if (condition) statement

for (i in range) statement

while (condition) statement

fun = function(input) {calculation return(output)}

fun = function(input) {calculation output}
```

Exercise 12 Programming

```
Write a program that asks the user to

type an integer N and compute u(N) defined with :

u(0)=1

u(1)=1

u(n+1)=u(n)+u(n-1)
```

1. Evaluate $1^2 + 2^2 + 3^2 + \ldots 400^2$.

Table 2: Apply functions

Functions	Usage
apply	Apply Functions Over Array Margins
by	Apply a Function to a Data Frame Split by Factors
eapply	Apply a Function Over Values in an Environment
lapply	Apply a Function over a List or Vector
mapply	Apply a Function to Multiple List or Vector Arguments
rapply	Recursively Apply a Function to a List
tapply	Apply a Function Over a Ragged Array

2. Evaluate $1 \times 2 + 2 \times 3 + 3 \times 4 + ... + 249 \times 250$

3. Create a function "crra" with two arguments (c, θ) that returns $\frac{c^{1-\theta}}{1-\theta}$. Add an if condition such that the utility is given by the log when $\theta \in [0.97, 1, 03] \approx 1$

4. Create a function "fact" that returns the factorial of a number

Exercise 13 Apply Functions

1. Using this object,

   ```
   m = matrix(c(rnorm(20,0,10), rnorm(20,-1,10)), nrow = 20, ncol = 2)
   ```

   Calculate the mean, median, min, max and standard deviation by row and column.

2. Using the dataset iris in the package "datasets", calculate the average **Sepal.Length** by **Species**. Evaluate the sum log of **Sepal.Width** by **Species**.

3.
   ```
   y1 = NULL; for (i in 1:100) y1[i]=exp(i)
   y2 = exp(1:100)
   y3 = sapply(1:100,exp)
   ```

   (a) Check the outcome of these three operations.

   (b) Using proc.time() or system.time(), compare the execution time of these three equivalents commands.

Table 3: Statistical distributions

name	description
dname( )	density or probability function
pname( )	cumulative density function
qname( )	quantile function
rname( )	random deviates

Table 4: Statistical Functions

Function	Description
mean(x, trim=0,na.rm=FALSE)	mean of object x
sd(x), var(x)	standard deviation, variance of object(x)
median(x)	median
quantile(x, probs)	x is the numeric vector and probs is a numeric vector with probabilities
range(x)	range
sum(x)	sum
diff(x, lag=1)	lagged differences, with lag indicating which lag to use
min(x)	minimum
max(x)	maximum

Table 5: Statistical distributions

Distribution	R name
Beta	beta
Lognormal	lnorm
Binomial	binom
Negative Binomial	nbinom
Cauchy	cauchy
Normal	norm
Chisquare	chisq
Poisson	pois
Exponential	exp
Student t	t
F	f
Uniform	unif
Gamma	gamma
Tukey	tukey
Geometric	geom
Weibull	weib
Hypergeometric	hyper
Wilcoxon	wilcox
Logistic	logis

1.5 Statistics

Exercise 14 Simulating and Computing

1. Simulate a vector x of 10,000 draws from a normal distribution. Use the function summary to provide basic characteristics of x.

2. Create a function *dsummary* that returns, the minimum, the 1st decile, the 1st quartile, the median, the mean, the standard deviation, the 3rd quartile, the 9th decile, and the maximum.

3. Suppose $X \sim N(2, 0.25)$. Evaluate $f(0.5), F(2.5), F^{-1}(0.95)$

4. Repeat if X has t-distribution with 5 degrees of freedom.

5. Suppose $X \sim P(3, 1)$, where P is the pareto distribution. Evaluate $f(0.5), F(2.5), F^{-1}(0.95)$

Exercise 15 Moments

Consider a vector $V = rnorm(100, -2, 5)$.

1. Evaluate n as the length of V.

2. Compute the mean $m = \frac{1}{n} \sum_{i=1}^{i=n} V_i$

3. Compute the variance $s^2 = \frac{1}{n-1} \sum_i^n (V_i - m)^2$

4. Compute the skewness $\gamma_1 = \frac{1}{n} \frac{(V_i - m)^3}{s^3}$

5. Compute the kurtosis $k_1 = \frac{1}{n} \frac{(V_i - m)^4}{s^4} - 3$

Exercise 16 OLS

1. Create a matrix X of dimension *(1000,10)*. Fill it with draws from a beta distribution with shape1 parameter 2, and shape 2 parameter 1. Make sure that there is no negative.

2. Create a scalar denoted by σ^2 and set it to 0.5. Generate a vector β of size 10. Fill it with draws from a *Gamma* distribution with parameters 2 and 1.

Table 6: Matrix operation

Function (Operator)	Description
$A * B$	Element wise multiplication
$A\% * \%B$	matrix multiplication
t(A)	Transpose
diag(a)	Create a diagonal matrix with a elements
diag(A)	Return the diagonal of A
Solve(A)	inverse of A

3. Create a vector ϵ of 1000 draws from a normal distribution.

4. Create $Y = X\beta + \sqrt{\sigma^2} * \epsilon$

5. Recover $\hat{\beta} = (X'X)^{-1}(X'Y)$

6. Evaluate $\widehat{\epsilon} = \widehat{y} - y$. Plot the histogram (filled in grey) and the kernel density of the distribution of the error term.

7. Estimate $\sigma^2 = \dfrac{\widehat{\epsilon}'\widehat{\epsilon}}{n-p-1}$, and $\mathbb{V}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$

8. Create *param* that binds $(\beta, \sqrt{V(\widehat{\beta})})$. Using the command *lm*, check these estimates.

9. Construct a confidence interval for β.

10. Redo the exercise by setting $\sigma^2 = 0.01$. How are your confidence intervals for β.