RAMNIRANJAN JHUNJHUNWALA AUTONOMOUS COLLEGE, GHATKOPAR (W)
Department of Statistics
T. Y. B. Sc. Semester- V Statistics Practical Paper IV

Practical No. 5.4.1 Fundamentals of R Date:

1. Create two vectors s1 and s2 containing marks obtained by 4 students and 8 students in two groups as given below:

   s1: 37, 49, 7, 38
   s2: 16, 37, 21, 42, 27, 40, 39, 51
   From vectors s1 and s2, create vectors, sp=s1+s2, sn=(s1+s2)/2, sd=s1/s2, sm=s1*s2.
2. Create a vector of number 1,4,7,….,37 and label as "a" and 1,2,………13 label as "b. Find no of observation in a and b , c=a*b, d=a/b, e=a+b, f=a-b

3. Following data gives ages of patients admitted to a hospital on a day
   40, 67, 75, 48, 44, 53, 56, 43, 66, 57, 65, 52, 83, 83, 80, 76, 85, 88, 89, 87
   Write a program for the following sub-questions:-
   a) Enter the data using c() and label as "y"

   b) Find mean, median, mode.

   c) Calculate log10, log and no. of observation and sort data in ascending order.

   d) Calculate summary of data, range, s.d.
4. Consider matrix A= and B =
   Write a program for the following sub-questions:-
   a) Enter the above matrices as given label.

   b) Calculate A+B, B-A, AB, IAI, IBI, inverse of A and B, A', AB', (AB)', A'B'.

5.

   Consider matrix A= and B =
   Write a program for the following sub-questions:-
   a) Enter the above matrices as given label.

   b) Calculate A+B, B-A, AB, IAI, IBI, inverse of A and B, A', AB', inverse(AB)', A'B'.

6.

Id no.	1	2	3	4	5	6	7	8	9	10

Write a

Id no.	1	2	3	4	5	6	7	8	9	10
Age of son(yrs)	5	3	2	6	2	3	7	4	5	7

program for the following sub-questions:
a) Prepare a data frame “d” as Id no. =n, Age of husband (yrs)=x, Age of wife (yrs)=y
b) Display only first 6 rows.
c) Display last 6 rows.
d) How many no of rows and columns.
e) Display only first 3 rows, last 3 rows, first 3 rows and two columns, 3rd row of 1st column entry
f) Create new variable Z=(x+y)/2 and add it to “d” as label z.

7.

Write a program for the following sub-questions:

a) Prepare a data frame “c” as Id no. =n, Age of son (yrs)=z1,

b) Append the variable z1 to data set “d” of question no. 6 and labeled as dataset “d1”.

c) Rename the variable z1 as z2 in data set “d1”

d) Delete variable “z2”, drop variable “x” and “y”.

8. Following are height (in cms) and weight (in kgs) of 10 boys.

Height: 140, 137, 150, 147, 139, 140, 150, 132, 138, 140

Weight: 55, 57, 59, 62, 61, 60, 60, 58, 59, 57

Write a program for the following sub-questions:

a) Prepare a data frame “d” of height and weight, label as height=”h” and weight=”w”.
b) Create a vector of boys with height> 145.
c) Create a vector of boys with weight> 55.
d) Create a data frame of the boys with height > 140 and weight > 60.

RAMNIRANJAN JHUNJHUNWALA AUTONOMOUS COLLEGE, GHATKOPAR (W), MUM-86

Department of Statistics

T. Y. B. Sc. Semester- V

Practical No.5.4.2 **GRAPHS AND DIAGRAMS** **Date:** **Roll No:**

Q.1Access stackloss data from R and answer the following:

a. Plot histogram and boxplot of Air flow.
b. Change the binwidth to 5 and also add appropriate labels and title to the plot
c. Plot Scatter plot of Air flow Vs Water temperature and also give labels, title, colour.

Q.2Using loan.xls , plot following graphs and diagrams

a. Simple bar plot of variable purpose
b. Multiple bar plot of Purpose of Loan and Creditability vs Count of Customers
c. Histogram of Credit Amount i) bin width=1000 ii) color="lightblue" with Title and labels
d. Boxplot of Credit Amount and colour it according to Creditability of customers with Title and labels.
e. Frequency polygon Credit Amount and also colour it according to Creditability of customers with Title and labels.
f. Scatter plot of Credit Amount v/s Age(years) and colour it according to Creditability of customers with Title and labels.

Q.3Following table gives the birth rate per thousand of different countries over the certain period:

Country	India	Germany	China	Pakistan	Sweden
Birth Rate	33	16	40	35	15

Represent the above data using Bar plot.

Q.4The following data represents maximum and minimum temperatures of the four metropolitan cities in India.

	Temperature in Celsius	
City	**Maximum**	**Minimum**
Delhi	40.5	34.7
Kolkata	42.8	33.5
Mumbai	37.8	32.2
Chennai	39.4	33.1

Represent the data using suitable diagram.

Q.5Draw frequency polygon for the data given below:

Marks	:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	:	6	12	25	16	11

Q.6During 2000 – 2001 to 2002 – 2003, the number of students in University 'X' (All figures are given in thousand) are as follows. Represent the data by subdivided bar diagram.

Year	**Arts**	**Science**	**Law**	**Total**
2000 – 2001	20	10	5	35
2001 – 2002	25	9	10	44
2002 – 2003	30	20	20	70

Q.7The following is the data regarding the average daily expenditure (in Rs.) of a family. Draw a Pie diagram to represent the data.

Items	Expenditur e
Food	240
Clothing	66
Rent	125
Fuel and Lighting	57
Education	42
Miscellaneous	190

Practice problems

Q.8 The following data referring to rainfall in a city of India in five years:

Year	2001	2002	2003	2004	2005
Rainfall (in cm)	195	228	186	205	246

Q.9 The following data refer to the birth rates and death rates of the following countries in 2007. Represent the information by appropriate diagram.

Country	Birth Rate	Death Rate
USA	14.2	8.3
China	17.7	6.2
Italy	8.5	10.5

Q.10 The following data represents the maximum and minimum temperature of the four metropolitan cities in India. Represent the data by suitable diagram.

City	Temp. In Celsius	
	Max.	Min.
Delhi	40.5	34.7
Kolkata	42.8	33.5
Mumbai	37.8	32.2
Chennai	39.4	33.1

@@@@@@@@@@@@@@@@@@ End @@@@@@@@@@@@@@@@@@@@@@@

RAMNIRANJAN JHUNJHUNWALA COLLEGE, GHATKOPAR (W), MUM-86 (AUTONOMOUS)

Department of Statistics

T. Y. B. Sc. Semester- V

Practical No.5.4.3 **Simple Linear Regression Model** **Date:** **Roll No:**

Q.1 Carry out the simple linear regression analysis and interpret the results for the following data sets whose detailed description is given below:

(i) Filename: Birdexctinct.xls

Problem: National Park size

Nature of data:

Column	Description
A	Site number
B	Area (Sq. Km)
C	# Species at risk
D	# Species extinct

Background: One of the major controversies in conservation biology is 'a few small versus many large'. The problem is that of optimum use of resources to conserve species. If we have limited land and we wish to use it to create protected areas for conserving say bird species, should we make one large national park out of it or should we have many small sanctuaries? This depends on extinction rates as a function of area of a park or sanctuary. If the relation is linear then it does not matter. If there are economies of scale, it may be better to have a few large parks. In a study of several islands in Finland, two surveys, one in 1949 and the other in 1959 were used to decide the number of species present and those that went extinct in 10 years. We need to check the relationship between the area and proportion that went extinct.

(ii) **Filename:** Crack.xls

Problem: Healing the heel

Nature of data:

Column	Description
A	Change in grade (severity) of cracking **(right heel)**
B	Change in typical crack length **(right heel)**
C	Change in grade (severity) of cracking (**left heel**)
D	Change in typical crack length (**left heel**)

Background: People who work bear foot often suffer from cracks in the heel. If the cracks are severe they can cause pain, bleeding, infection etc. Many traditional remedies are in use for this ailment. In a study to test efficacy of an ayurvedic treatment, severity of cracking was recorded and also typical length of a crack. This was done for each heel before and after treatment.

(iii)**Filename:** crime.xls

Problem: Relation between crime and intelligence

Nature of data :

Column	Description
A	delinquency index
B	IQ

Background: It is of interest to know the relationship between intelligence of the criminal and his delinquency (crime) index (from 0 to 50), which is a combination of frequency of crime and seriousness of criminal acts of an individual. This may help in 'managing' the case in jail. So we need to know the general rule and exceptions if any etc.

Prepare a report on the nature of relationship between the two variables. It should include essential technical details and should guide a non-statistician who has to use it in his job of jail management.

(iv)**File name**: moth.xls

Problem: Natural selection

Nature of Data:

Column	Description
A	**site number**
B	**Distance from city**
C	**Moth type (light / dark)**
D	**# moths placed**
E	**# moths removed by predators**

Background: The basic principle of Darwin's theory of evolution through natural selection is that as environment changes, ability of an organism to survive also changes. This was experimentally tested in and around Liverpool in United Kingdom. A moth species that comes in two varieties (light and dark) was used. Trees in Liverpool have blackened trunks due to industrial smoke. The darkness reduces as we go farther from the city. Dark moths can blend with the dark trunks and hence rate of predation is lower for this variety in the vicinity of Liverpool. As the distance of a locality from Liverpool increases and tree trunks become lighter, pendulum shifts in favour of the light variety. In the experiment in question, dead moths were left on tree trunks and were revisited after 24 hours. The number of moths removed (presumably by predators) was recorded.

RAMNIRANJAN JHUNJHUNWALA COLLEGE, GHATKOPAR (W), MUM-86 (AUTONOMOUS)

Department of Statistics

T. Y. B. Sc. Semester- V

Practical No.5.4.4 **Multiple Linear Regression Model** **Date:** **Roll No:**

Q.1 Carry out multiple linear regression analysis for the following data sets whose description is given below:

(i) Filename: mammalsize.xls

Problem: Correlates of brain size

Nature of data :

Column	Description
A	name of the species
B	gestation period (days)
C	brain weight (gms)
D	body weight (kg)
E	litter size

Background: Data are from American Naturalist (1974) p.593-613. Animals have properties that make them better capable of living and multiplying. One expects that larger brain may be generally better. But there can be penalties and limitations. One limitation is need for longer pregnancy and the other is the need to have fewer offsprings.

(ii) **Filename:** Crack_combined.xls

Problem: Healing the heel

Nature of data:

Column	Description
A	Change in grade (severity) of cracking
B	Change in typical crack length
C	Type of leg. 1 represents left leg while 0 represents right leg.

Background: People who work bear foot often suffer from cracks in the heel. If the cracks are severe they can cause pain, bleeding, infection etc. Many traditional remedies are in use for this ailment. In a study to test efficacy of an ayurvedic treatment, severity of cracking was recorded and also typical length of a crack.

(iii)**File name**: moth.xls

Problem: Natural selection

Nature of Data:

Column	Description
A	**site number**
B	**Distance from city**
C	**# moths placed**
D	**# moths removed by predators**
E	**Colour of moth. 1 represents dark coloured while 0 represents light-coloured moths.**

Background: The basic principle of Darwin's theory of evolution through natural selection is that as environment changes, ability of an organism to survive also changes. This was experimentally tested in and around Liverpool in United Kingdom. A moth species that comes in two varieties (light and dark) was used. Trees in Liverpool have blackened trunks due to industrial smoke. The darkness reduces as we go farther from the city. Dark moths can blend with the dark trunks and hence rate of predation is lower for this variety in the vicinity of Liverpool. As the distance of a locality from Liverpool increases and tree trunks become lighter, pendulum shifts in favor of the light variety. In the experiment in question, dead moths were left on tree trunks and were revisited after 24 hours. The number of moths removed (presumably by predators) was recorded.

(iv)**File name**: Autompg.xls

Problem: To Identify which variables are influencing the miles per gallon(MPG i.e fuel consumption) of a car

Nature of Data:

Variables	**Variable Type**
mpg(Miles per gallon)	continuous
cylinders	multi-valued discrete
displacement	continuous
horsepower	continuous
weight	continuous
acceleration	continuous
model year	multi-valued discrete
origin	multi-valued discrete
car name	string (unique for each instance)

Background: The data consists of 398 records characterizing various car types. For each car type the following attributes are provided: the MPG value (mpg) measured for each car model in a test performed in 1982, the number of the engine cylinders (cyl), the cylinder displacement in cubic inches (displ), the engine power (power), the car weight in pounds (weight), a number of seconds required to accelerate to the speed of 100 miles per hour (accel), the car's production year (year), the country of production (origin: USA, Europe, or Japan), and the name of the model (model).

Moth

RAMNIRANJAN JHUNJHUNWALA COLLEGE, GHATKOPAR (W), MUM-86 (AUTONOMOUS)

Department of Statistics

T. Y. B. Sc. Semester- V

Practical No.5.4.5 Multicollinearity and Weighted Least Square Date: Roll No:

Q.1 Following dataset spans the period from 1947 to 1966 and presents observation on the variables: imports (I), Gross domestic Product(GDP), Stock information (SF) and Consumption (C). Check the multicollinearity and use remedial measures.

Import	GDP	SF	consumption
15.9	149.3	4.2	108.1
16.4	161.2	4.1	114.8
19	171.5	3.1	123.2
19.1	175.5	3.1	126.9
18.8	180.8	1.1	132.1
20.4	190.7	2.2	137.7
22.7	202.1	2.1	146
26.5	212.4	5.6	154.1
28.1	226.1	5	162.3
27.6	231.9	5.1	164.3
26.3	239	0.7	167.6
31.1	258	5.6	176.8
33.3	269.8	3.9	186.6
37	288.4	3.1	199.7
43.3	304.5	4.6	213.9
49	323.4	7	223.8
50.3	336.8	1.2	232
56.6	353.9	4.5	242.9

Q.2 Check the multicollinearity and use remedial measures for the data whose description is given below:

File name: 1 auto.xls

Problem: To Identify which variables are influencing the miles per gallon(MPG i.e fuel consumption) of a car

Nature of Data:

Variables	**Variable Type**
mpg(Miles per gallon)	Continuous
Cylinders	multi-valued discrete
Displacement	Continuous
Horsepower	Continuous
Weight	Continuous
Acceleration	Continuous
model year	multi-valued discrete
car name	string (unique for each instance)

Background: The data consists of 398 records characterizing various car types. For each car type the following attributes are provided: the MPG value (mpg) measured for each car model in a test performed in 1982, the number of the engine cylinders (cyl), the cylinder displacement in cubic inches (displ), the engine power (power), the car weight in pounds (weight), a number of seconds required to accelerate to the speed of 100 miles per hour (accel), the car's production year (year) , and the name of the model (model).

Q.3 In a drug stability evaluation, an antimicrobial product was held at ambient temperature (~68^{o}F) for 12 months. The potency (%) through HPLC was measured, 10^6 colony-forming units (CFU) of *Staphylococcus aureus*(methicillin-resistant) were

exposed to the product for 2 min, and the microbial reductions ($\log_{10}$ scale) were measured. The following table provides the data and weights. Using this data, fit a WLS model.

Y	x1	x2	wts
100	1	5	1
100	1	5	1
100	1	5.1	1
100	2	5	1

100	2	5.1	1
100	2	5	1
98	3	4.8	0.45
99	3	4.9	0.45
99	3	4.8	0.45
97	4	4.6	0.23
96	4	4.7	0.23
95	4	4.6	0.23
95	5	4.7	0.07
87	5	4.3	0.07
93	5	4.4	0.07
90	6	4	0.03
85	6	4.4	0.03
82	6	4.6	0.03
88	7	4.5	0.24
84	7	3.2	0.24
88	7	4.1	0.24
87	8	4.4	0.1
83	8	4.5	0.1
79	8	3.6	0.1
73	9	4	0.03
86	9	3.2	0.03
80	9	3	0.03
86	10	4.2	0.04
83	10	3.1	0.04
72	10	2.9	0.04
70	11	2.3	0.03
88	11	3.1	0.03
68	11	1	0.03
70	12	1	0.03

68	12	2.1	0.03
52	12	0.3	0.03

y = potency, the measure of the kill of *Staphylococcus aureus* following a 2 min exposure;

100% = fresh product = 5 $\log_{10}$ reduction.

x_1= month of test = end of month.

x_2 = $\log_{10}$ reduction in a 10^6 CFU population of *S.aureus*in 2 min.

wts = weights

Q.4 Estimate the model D = b_0 + b_1P + u using WLS method. The data is given below:

D	P	Var(u)
10	8	1
20	7	4
23	6	5
25	4	2
42	3	1