MACHAKOS UNIVERSITY

University Examinations 2021/2022 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR FIRST SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (STATISTICS AND PROGRAMMING)

SST 301 PROGRAMMING LANGUAGE FOR STATISTICS I

DATE: 30/8/2022 **TIME: 2.00-4.00 PM**

INSTRUCTION:

Answer Question One and Any Other Two Questions

QUESTION ONE (30 MARKS)

a) Given the following vectors, `vec1` $< -c(1,2,3,4,5)$, `vec2` $< -c(1,3,45,67,89,32)$, write R output for the following functions

i. `union(vec1,vec2)`

ii. `setdiff(vec1,vec2)`

iii. `setequal(vec1,vec2)` (3 marks)

b) Write R output for the following

i. `matrix(c(1,2,3,4,5,6),nrow=2,ncol=3,byrow=TRUE)` (2 marks)

ii. `rep(x=c(3,62,8.3), times=3)` (2 marks)

iii. `rep(c(3,62,8.3), each=2)` (2 marks)

iv. `factor(c(1,0,1,0,0,0),levels=c(0,1),labels=c("male", "female"))` (2 marks)

c) Let $A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$. Write R code to determine the inverse of A. (3 marks)

d) Given the table below;

Name	**Age**	**Sex**
Peter	42	M
Lois	40	F
Meg	17	F
Chris	14	M
Stewie	1	M

i. Create a data frame called *mydata* with the data in R, defining the *Sex* variable as a factor. (5 marks)

ii. Write an R command to extract the third, fourth and fifth elements of the third column. (1 mark)

iii. Write an R command to extract the entire third and first columns. (1 mark)

iv. Write R command to add the variable *funny* defined as *funny*=c('High', 'High', 'Low', 'Med', 'High', 'Med') to the data frame (2 marks)

v. Write R code to extract all rows where $Age > 10$ OR degree of funniness is 'High'. (2 marks)

e) Write a function in R with two input variables which are numerical. The function should add, subtract, multiply and divide the two variables and print as output all these. Further, an error message should be displayed if the input variable is not numerical or is zero.

(5 marks)

QUESTION TWO (20 MARKS)

a) Write down the simple linear regression model explaining all its components and the assumptions taken into account when defining the model. (6 marks)

b) Data was collected from 392 vehicles on several variables such as gas mileage, horsepower, engine displacement, vehicle weight among others. Assume that the data is in a notepad file named *autodata* stored in directory F and folder SST301 in your computer. Write R code that does the following:

i. Reads in the data into R. (2 marks)

ii. Does a scatter plot of the variables *mpg* (response) and *horsepower* (predictor) with labels MPG and Horsepower for the Y and X axes and has the title 'SCATTER PLOT OF AUTO DATA' (3 marks)

iii. Fits a simple linear regression model. (2 marks)

iv. Superimposes the fitted regression line on the scatter plot. (1 mark)

c) A simple linear regression model was fit to the data above and the following output was obtained:

Residuals:

Min	1Q	Median	3Q	Max
-13.5710	-3.2592	-0.3435	2.7630	16.9240

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.935861	0.717499	55.66	<2e-16 ***
horsepower	-0.157845	0.006446	-24.49	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom

Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049

F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

i. From the output above, comment on the following

I. Is there a relationship between the predictor and the response? (1 mark)

II. How strong is the relationship between the predictor and the response? (1 mark)

ii. Determine the 95% confidence intervals for the intercept and horsepower. (4 marks)

QUESTION THREE (20 MARKS)

a) State two assumptions underlying an analysis of variance. (2 marks)

b) An investigation was carried out into the number of hours (*hours*) lost due to stress-related illnesses at various companies (*company*). The null hypothesis stated that the means for the five companies sampled were equal while the alternative hypothesis stated that not all five companies' means are equal. ANOVA was used to test this hypothesis and the R output is as below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
company	4	6507	1626.7	12.78	0.000138 ***
Residuals	14	1782	127.3		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

i. Write the R command that produced the output above. (2 marks)

ii. Interpret the output and give your conclusion for the hypothesis under test. (3 marks)

iii. Explain what can be done to determine which groups are different from each other. Write R code that can be used to test this. (2 marks)

c) A study was carried out to assess the value of houses in a certain neighbourhood. A multiple regression was carried out to determine whether the variables *age* (average age of houses) and *lstat* (percent of houses with low socioeconomic status) had an effect on *medv* (median house value).

i. Write the R code for this. (1 mark)

ii. The following output was obtained

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,  Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

Write down the regression line. (1 mark)

iii. Write down the hypothesis that you would test to check whether the explanatory variables are contributing significantly in predicting the response variable.

(2 marks)

iv. From the output above, what would be the result of your tests in iii above at 5% significant level? (3 marks)

v. What is the difference between the Multiple R squared and Adjusted R squared as given in R output? (2 marks)

vi. According to the output, would you say that the model is a good fit? Explain your answer. (2 marks)

QUESTION FOUR (20 MARKS)

a) i. Construct a 4 x 2 matrix that's filled row wise with the values 4.3, 3.1, 8.2, 8.2, 3.2, 0.9, 1.6 and 6.5 in that order. Store it under the variable name *newmat*. (2 marks)

ii. Overwrite the second column of the matrix from (i) with that same column sorted from smallest to largest. (2 marks)

b) Create a list that contains, in this order, a sequence of 20 evenly spaced numbers between -4 and 4; a 3 x 3 matrix of the logical vector c(F,T,T,T,F,T,T,F,F) filled column-wise; a character vector with the two strings "don" and "quixote"; and a factor vector containing the observations c ('Low', 'Med', 'Low', 'Med', 'Med', 'High'). (8 marks)

c) A study was conducted to investigate the relationship between the price of a plot of land and the distance from the nearest tarmac road. A random sample of 12 plots was taken and the data relating the distance from the nearest tarmac road in hundred metres and the price of the plot in million Kenya shillings is as shown in the table below.

Plot	1	2	3	4	5	6	7	8	9	10	11	12
Distance	48	78	35	72	10	60	40	22	84	25	12	48

Price	1.2	0.4	1.4	0.8	1.8	0.6	1.5	1.2	0.8	1.5	1.6	1.2

i. Write R code to calculate the mean and standard deviation of the variable *Price* (1 mark)

ii. Write R code to compute the Pearson's product moment co-efficient of correlation between the distance from the tarmac and the price of the plot. (2 marks)

d) Six people are asked to rate the performance of a product on a scale of 1-5, with 1 representing very poor performance and 5 representing very good performance. The following data on their responses 1,3,4,2,2,1 was obtained. Write R code to represent this data and give the expected output. (5 marks)

QUESTION FIVE (20 MARKS)

a) Outline the procedure used for hypothesis testing. (4 marks)

b) Differentiate between Type I and Type II errors as used in hypothesis testing. (2 marks)

c) In order to increase the efficiency with which employees in a certain organisation can carry out a task, 5 employees are sent on a training course. The time in seconds to carry out the task both *before* and *after* the training course for 5 employees was recorded.

i. Assuming the data is already saved in the workspace as *employeedata*, with the variables *before* and *after*, write R code to test whether the training course has had the desired effect. (2 marks)

ii. Assuming you run the code in R and part of the results are as given below:

```
t = 1.6842, df = 4, p-value = 0.08372
```

State and give the interpretation of your hypothesis, taking $\alpha = 0.05$. (3 marks)

d) The table below shows the numbers of births during one month at a particular hospital classified according to whether a particular medical characteristic was or wasn't present during childbirth.

Age of mother	**<20**	**21-25**	**26-30**	**31-35**	**36+**	**Total**
Characteristic present	10	12	9	4	3	38
Characteristic absent	5	51	38	25	5	124
Total	15	63	47	29	8	162

i. Determine whether the presence of this characteristic is dependent on the age o f the mother. (7 marks)

ii. Write R code that could have assisted you produce the same results as above, a ssuming the file is already in R under the name *birthdata*. (2 marks)