Tutorial letter 013/0/2023

Applied Statistics III

STA3701

Year module

Department of Statistics

ASSIGNMENT 03 QUESTIONS

Define tomorrow.

UNISA
university of south africa

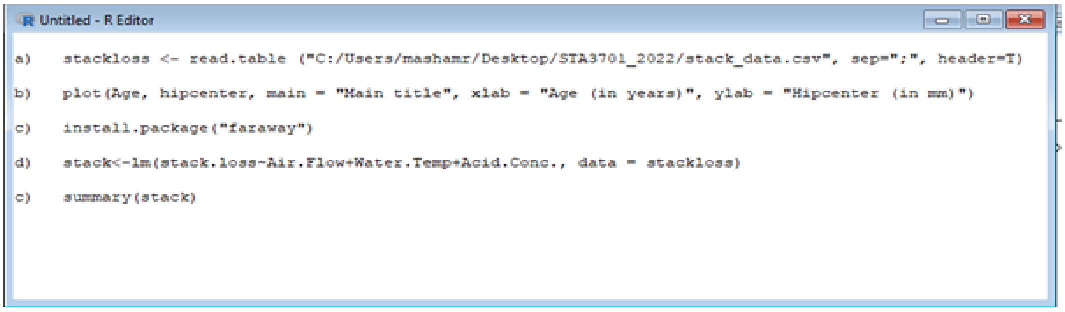
ASSIGNMENT 03
Unique Nr.: 859703
Fixed closing date: 7 August 2023

Instructions

1. **Do not PLAGIARISE. Students suspected of plagiarism will be subjected to disciplinary processes.**
2. **Use R to answer all the questions. Present or attach R outputs. Label all the figures and tables.**
3. **For all the hypothesis tests, use a level of significance of five percent.**

QUESTION 1 [10]

Study the R syntax displayed on Figure 1 and answer the questions that follow:



```
a)   stackloss <- read.table ("C:/Users/mashamr/Desktop/STA3701_2022/stack_data.csv", sep=";", header=T)

b)   plot(Age, hipcenter, main = "Main title", xlab = "Age (in years)", ylab = "Hipcenter (in mm)")

c)   install.package("faraway")

d)   stack<-lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc., data = stackloss)

c)   summary(stack)
```

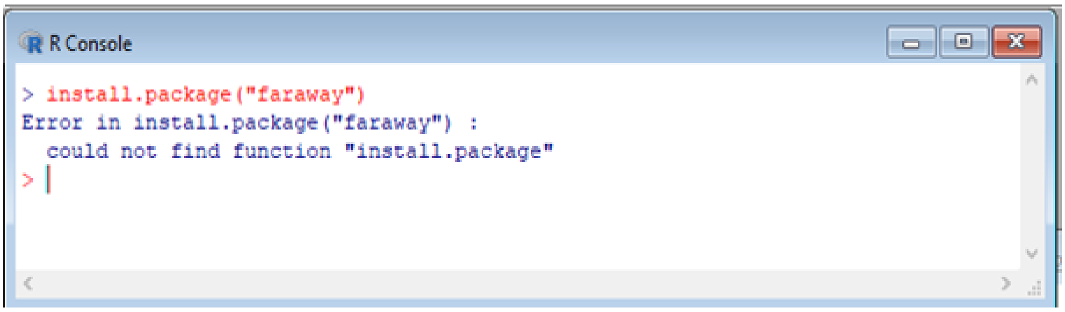
Figure 1: R Syntax

1.1 Describe the function of the commands:

- 1.1 read.table () (1)
- 1.2 plot () (1)
- 1.3 lm () (1)
- 1.4 summary () (1)

1.2 Refer to code (b). What type of plot will the command produce? Specify the independent variable (x) and dependent variable (y) in the code. (3)

1.3 Running code (c) gives the feedback shown on figure 2 below.



```
> install.package("faraway")
Error in install.package("faraway") :
  could not find function "install.package"
>
```

Figure 2: Error message feedback from running code (c)

Give the correct command and briefly explain the function of the command. (3)

QUESTION 2 **[35]**

Consider the dataset USArrests that is an inbuilt dataset in R then answer the questions below.

2.1 Which R package is the data found in? (1)

2.2 Give a brief description of the dataset specifying the variables that were observed in the study. (4)

2.3 Suppose that your aim is to investigate the relationship between urban population percentage (x-variable) and murder (y-variable):

i) Make a scatterplot of the two variables and comment on their relationship. (5)

ii) Fit a simple linear regression to model the relationship. Give the fitted model (regression equation) and interpret the coefficients. (5)

iii) Perform tests of hypotheses on the regression coefficients and comment on their significance. For all the tests, provide the null and alternative hypotheses, critical region (or rejection region), test statistics and your conclusions. (8)

iv) Use relevant diagnostic plots and/or hypothesis tests to investigate if the assumptions of the model were violated. (12)

QUESTION 3 **[70]**

Pima.tr dataset in R contains a sample of 200 Pima Indian women living near Phoenix, Arizona. They were tested for diabetes according to World Health Organization criteria. The US National Institute of Diabetes and Digestive and Kidney Diseases collected the data. Table 1 gives the name and description of each variable observed in the study.

Table 1: Variables observed in the Pima.tr (Diabetes in Pima Indian Women)

Variable abbreviaton/name	Description
npreg	Number of pregnancies
glu	Plasma glucose concentration in an oral glucose tolerance test
bp	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
bmi	Body mass index (weight in kg/(height in m^2))
ped	Diabetes pedigree function
age	Age in years
type	Yes/no for diabetes according to WHO criteria

Fit a multiple regression model to the Pima.tr data with bp as the dependent variable and all the variables excluding type as independent variables, then answer the questions that follow.

3.1 Are the assumptions of homogeneous variance and normality of residuals violated? Explain. (8)

3.2 Check the model for leverage points, outliers, and influential points. Discuss your results. (15)

3.3 Use a scatterplot matrix and correlation matrix to examine the existence of collinearity between the predictors in the model. Report your findings. (10)

3.4 Compute variance inflation factors. Do the results corroborate your observation in 3.3? (7)

3.5 Refer to the findings in part 3.1, can you recommend that the response variable be transformed? Use the Box–Cox method to determine the best transformation on the response. (10)

3.6 Select the “best” model by using the Akaike information criterion (AIC). Give the fitted model and interpret the coefficients. (8)

3.7 Use the p-value approach to test the significance of the regression coefficients. For all the tests, provide the null and alternative hypotheses, critical region (or rejection region) and your conclusions. (12)

Grand total: [115 marks]