MANICALAND STATE UNIVERSITY OF APPLIED SCIENCES

FACULTY OF APPLIED SCIENCES & TECHNOLOGY

DEPARTMENT OF APPLIED STATISTICS

MODULE: STATISTICAL COMPUTING I

CODE: HAST215

SESSIONAL EXAMINATIONS

DECEMBER 2023

EXAMINER: MRS S MANDIZVIDZA

INSTRUCTIONS

1. Answer **All** in Section A.
2. Answer **three** questions in Section B.
3. Start a new question on a fresh page.
4. Total marks: 100.

Additional material(s)

- Statistical tables, Non-programmable electronic scientific calculator, List of formulae.

SECTION A [40 MARKS]

Answer **ALL** questions in this section

A 1

Write the R codes for finding the following

(a) Number of variables in a dataset. (2)

(b) Variance. (2)

(c) Mean. (2)

(d) Minimum. (2)

A 2

(a) State any three assumptions of simple linear regression analysis. (3)

(b) State and explain how you carry out residual tests and linear regression assumptions using each of the following. (6)

1. Scatter plots
2. Normal probability plot (P-P plots).
3. Quantile Plots (Q-QPlots).

A 3

Write an R function to fit a simple linear regression with or without intercept. The function should start linear.regression <- function(x, y, intercept=TRUE)

(a) Have the function return the coefficients, the fitted values, the residuals, and the residual sum-of-squares. (4)

(b) Use the QR-decomposition to perform the regression. (2)

A 4

Give commands used in R software to carry out basic inference on

(a) Student's t-test (2)

(b) F test to compare two variances (2)

(c) Pearson's chi squared test for count data (2)

(d) test for association (2)

A 5

(a) State simple linear regression assumptions. (3)

(b) Describe and explain how these assumptions are evaluated. Describe how the evaluation can be done using R software. Also use diagrams as aid to your explanations. (6)

SECTION B [60 MARKS]

Answer any **THREE** questions in this section

A 6

(a) Differentiate between the two R help commands *help ("keyword")* and *help.search ("keyword")* (2)

(b) State any three types of R objects. (3)

(c) Describe how data is imported from STATA into R. (3)

(d) Describe how data is imported from Excel into R. (4)

(e) Give the R command for finding the inverse of a matrix A. What command would you type in R to save your workspace? (2)

(f) What command would you type in R to save load a package? (2)

(g) What command would you type in R to get summary statistics for all the variables in a dataframe (2)

(h) Define a data frame in R. (2)

A 7

(a) Define correlation coefficient. (2)

(b) What are the uses of the correlation matrix when doing multiple linear regression modelling? (2)

(c) Describe how you would test for the significance of the correlation coefficient. (4)

(d) What are the advantages of using a stem and leaf diagram compared to a box and whisker plot? (5)

(e) Describe how one would use stepwise regression technique in multiple linear regression analysis. Describe the processes involved until one produces a parsimonious model. (7)

A 8

The analysis output for MPG dataset.

Model	β	std error	t- value	p-value
constant	5275.160	6037.505	.874	.385
mpg	-22.762	72.139	-.316	.753
weight	5.918	1.023	5.782	.000
length	-78.783	35.093	-2.245	.028
turn	-149.701	116.549	-1.284	.203
car market	3273.408	687.059	4.764	.000

Figure 1: MPG data

(a) From the analysis output, which independent variables are statistically significant and which ones are not. (3)

(b) Give a practical meaning of the regression coefficients for the variables weight and mpg in explaining the price of a motor vehicle. (4)

(c) Describe the other method not shown in the table for testing the significance of the independent variables in multiple regression modeling. Explaining how the method is used to judge the significance of the independent variable. (5)

(d) Describe and explain how goodness of fit of the overall regression model is conducted. Stating the statistics that can be used in the assessment. Also state the hypotheses to be tested in each case. (8)

A 9

The speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop. The simple linear regression model is defined by, $Y_i = \beta_0 + \beta_1 X_i + e_0$ where $e_0 \sim N(0, \sigma^2)$ for the car speet and its stopping distance.

```
> stop_dist_model = lm(dist ~ speed, data = cars)
> summary(stop_dist_model)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Figure 2: Car stopping distance

(a) Fit the linear regression model. (2)

(b) Comment on the significance of of the model coefficients. (5)

(c) Predict the stop-dist-model, when speed $= 8$. (1)

(d) What does the residual standard error tells you. (2)

(e) Carry out a significance test on the linear relationship between speed and stopping distance. (5)

(f) From the above table, which statistics are used to test the importance of a variable in a multiple regression model? Explain how the statistics are used to judge the significance of the independent variables. (5)