MANICALAND STATE UNIVERSITY OF APPLIED SCIENCES

FACULTY OF ENGINEERING APPLIED SCIENCES AND TECHNOLOGY

DEPARTMENT: APPLIED STATISTICS

MODULE: STATISTICAL COMPUTING II

CODE: ASTA 215

SESSIONAL EXAMINATIONS
APRIL 2024

DURATION: 3 HOURS

EXAMINER: MR. S. CHAMUNORWA

INSTRUCTIONS

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

Additional material(s): *Non-programmable electronic scientific calculator. Statistical tables*

SECTION A: 40 MARKS.
Attempt/Answer All

A1 a) What commands can be used to import dataset from Excel format and inspect the data, including its attributes and variable values?

b) Explain the role of do files in automating data management tasks.

[5,5]

A2 a) In statistical analysis what is the difference between correlation and regression.

b) Discuss the advantages of using STATA for statistical analysis.

[5,5]

A3 a) Give examples of real life situations where data is skewed.

b) Measures of dispersion (spread) which include the standard deviation(s) and the coefficient of variation (CV) can be used to make some inferences in financial decisions. Between the two measures, which on is more reliable and why? Rates of return over the past 6 years for two mutual funds are shown in 1.

Table 1: Fund Performance

Fund A	8.3	-6.0	18.9	-5.7	23.6	20
Fund B	12	-4.8	6.4	10.2	25.3	1.4

Which fund has a higher risk?

[4,6]

A4 a) Identify Stata commands utilized for conducting descriptive statistics.

[10]

SECTION B: [60 MARKS].
Attempt/Answer any 3

B5 Given the following multiple linear regression equation:

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \beta_3\mathbf{X_3} + \epsilon$$

a) Explain how you would estimate parameters in the regression equation using the Least Square method?

b) Calls to the New York Auto Club are possibly related to the weather, with more calls occurring during bad weather. This example illustrates descriptive analyses and simple linear regression to explore this hypothesis in a data set containing information on calendar day, weather (low temperatures (low)) and numbers of calls (calls). Use the regression output in Table 3 to answer the following questions

Table 2: Regression results

```
. regress calls low
```

Source	SS	df	MS		
Model	100233719	1	100233719	Number of obs	= 28
Residual	95513596.2	26	3673599.85	F(1, 26)	= 27.28
				Prob > F	= 0.0000
Total	195747315	27	7249900.56	R-squared	= 0.5121
				Adj R-squared	= 0.4933
				Root MSE	= 1916.7

calls	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
low	-145.154	27.78868	-5.22	0.000	-202.2744	-88.03352
_cons	7475.849	704.6304	10.61	0.000	6027.46	8924.237

Table 3: multiple linear regression results output

i) How much variance in calls is explained by low? Is this variance explained significantly different to 0?

ii) What is the slope? Is the slope statistically significant?

iii) Write out the model regression equation.

iv) If the weather has a low temperature of 15 degrees, what would be the predicted calls per day?

v) What would be the approximate 95% confidence interval of our prediction?

c) Stata provides four graphical approaches for evaluating a model as shown in Figure 1:

i) Evaluate the suitability of linear regression in explaining the data.

[4,3,2,1,2,3,5]

B6 a) A psychology student, Sarah, has distributed sleep diaries to her university friends to monitor the number of hours of sleep

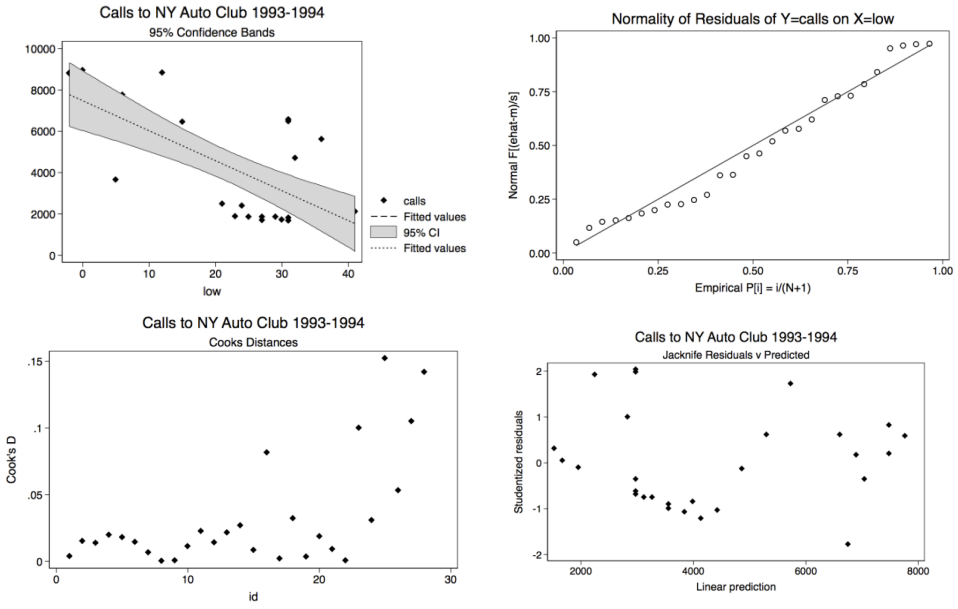

Figure 1:

they have each night. Sarah believes that university students, on average, sleep for six hours per night (HSN). The number of hours of sleep per night for each student was averaged over a one-month monitoring period. Table 2 displays the STATA output.

Figure 2: T-tets results

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
HSN	10	6.31	.3819395	1.207799	5.445993	7.174007

mean = mean(**HSN**) t = **0.8116**
Ho: mean = **6** degrees of freedom = **9**

Ha: mean < **6** Ha: mean != **6** Ha: mean > **6**
Pr(T < t) = **0.7810** Pr(|T| > |t|) = **0.4379** Pr(T > t) = **0.2190**

i) Write a code to perform one-sample t-test using the provided sleep duration data in the variable HSN.

ii) Is there any evidence to suggest that Sarah's belief is incorrect?

b) A research team investigated the usefulness of relaxation training

for reducing anxiety levels in individuals experiencing high-stress jobs. Thirty people were randomly selected from a group of 100 with high-stress jobs and divided into two groups. One group served as the control group (no training), and the other received relaxation training. A colleague repeats the experiment, matching samples on the dimensions of sex and job type. The results for the matched pairs are given in Table 4

Table 4: Paired t-tets results

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Control	15	30	1.712698	6.63325	26.32663	33.67337
Relaxa~n	15	24.73333	2.251913	8.721621	19.90346	29.56321
diff	15	5.266667	1.224615	4.742915	2.640128	7.893205

mean(diff) = mean(Control - Relaxation) t = 4.3007
Ho: mean(diff) = 0 degrees of freedom = 14

Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0
Pr(T < t) = 0.9996 Pr(|T| > |t|) = 0.0007 Pr(T > t) = 0.0004

i) Write a code to perform matched sample t-test.
ii) Evaluate her experiment using the criteria of $p < 0.05$ for a two-tailed test.

[10,10]

B7 a) Provide STATA commands to
i) change the order of observations based on the *name* variable.
ii) modify the name of the variable *age* to *age2*
b) Using examples, demonstrate how to recode values in the *age* variable.
c) Explain the purpose of data labels, variable labels and value labels in STATA.
d) Add a data label and a variable label to the dataset.
e) Create a new variable *age4* that is the result of adding *age2* and *age3*.

[3,4,3,3,2,5]

B8 Sudden death is an important, lethal cardiovascular endpoint. Most previous studies of risk factors for sudden death have focused on

men. Looking at this issue for women is important as well. For this purpose, data were used from the Framingham Heart Study. Several potential risk factors, such as age, blood pressure and cigarette smoking are of interest and need to be controlled for simultaneously. Therefore a multiple logistic regression was fitted to these data as shown in Table 5. The response is 2-year incidence of sudden death in females without prior coronary heart disease.

Table 5: multiple logistic regression results

Risk Factor	Regression Coefficient	Standard Error	Z	$P-value$
Constant	-15.3			
Blood Pressure (mm Hg)	.0019	.0070	.27	.7871
Weight (% of study mean)	-.0060	.0100	-.60	.5485
Cholesterol (mg/100 mL)	.0056	.0029	1.93	.0536
Glucose (mg/100 mL)	.0066	.0038	1.74	.0819
Smoking (cigarettes/day)	.0069	.0199	.35	.7623
Hematocrit (%)	.111	.049	2.27	.0235
Vital capacity (centiliters)	-.0098	.0036	-2.72	.0065
Age (years)	.0686	.0225	3.05	.0023

a) Assess the statistical significance of the individual risk factors and explain the practical implications of your findings.

b) Give brief interpretations of the age and vital capacity coefficients.

c) Predict the probability of sudden death for a 50 year old woman with systolic blood pressure of 120 mmHg, a relative weight of 100%, a cholesterol level of 250 mg/100mL, a glucose level of 100 mg/100mL, a hematocrit of 40%, and a vital capacity of 450 centiliters who smokes 10 cigarettes per day.

[3,7,10]

END OF QUESTION PAPER