MANICALAND STATE UNIVERSITY OF APPLIED SCIENCES

FACULTY OF ENGINEERING, APPLIED SCIENCES & TECHNOLOGY

DEPARTMENT OF APPLIED STATISTICS
LINEAR MODELS

CODE: ASTA 411

SESSIONAL EXAMINATIONS
2024

DURATION: 3 HOURS

EXAMINER: MR A. CHAKAIPA

INSTRUCTIONS

1. *Answer* ***All questions*** *in Section A*
2. *Answer any* ***two*** *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: Answer all questions [40 MARKS]

A1

a) Describe how one performs lacks of fit test to assess model adequacy

b) Model diagnostics in linear models involve examination of residuals (both graphically and using hypothesis testing techniques).

i) State any three graphical residual techniques employed, uses and any results expected.

ii) State any three hypotheses testing techniques, uses and any results expected.

[3, 9, 9]

A2

a) Define the following terms used in Analysis of Variance of Experimental Designs:

i) Single-factor experiment

ii) Random -effects experiment

b) Describe remedial transformation measures to employ to correct for each of the following:

i) non-linearity of the regression function

ii) Non-constancy of error variance **[2, 2, 2, 3]**

A3

a) In model diagnostic tests transformations are performed on either X or Y variable. State any advantages and or challenges met in trying to perform such transformations.

b) Explain why the coefficient of determination statistic may be misleading as a measure of model adequacy. **[6, 4]**

SECTION B: Answer any **two (2)** questions [**60 MARKS**]

B4

a) i) State any two situations where general linear models may not be appropriate to apply them.

ii) Define the following:

1. Generalized Linear model (GLM);
2. Link function.

iii) State any three assumptions of Generalized Linear Models.

iv) Explain any three advantages of GLMs over traditional ordinary Least Squares Regression (OLS).

b) Consider a simple logistic regression model output (based on the 2006 Health and Retirement Study [HRS] data) of the probability that a U.S. adult age 50+ has arthritis. The predictors in this main effects only model are gender, education level
[with levels less than high school (<12 years), high school (12 years), and more than high school (>12 years)], and age. The dependent variable is a respondent being diagnosed or not with arthritis. The results are summarized in Table 1.

i) Assuming a logit binomial model can employed to model the data, write a STATA code to model the data.

ii) Interpret the output in Table 1 for each of the explanatory variables, including the interaction term.
(Hint make use of odds ratios in your interpretation)

[3, 2, 2, 4, 4, 3, 12]

Table 1: Health and Retirement Study (HRS) on Arthritis data

Estimated Logistic Regression Model for Arthritis, Including the First-Order Interaction of Education and Gender

Predictor[a]	Category	$\hat{B}$	$se(\hat{B})$	t	$P(t_{56} > t)$
INTERCEPT	Constant	–2.728	0.135	–20.22	< 0.01
GENDER	Male	–0.659	0.061	–10.81	< 0.01
ED3CAT	<12 yrs	0.454	0.063	7.20	< 0.01
	12 yrs	0.177	0.050	3.56	< 0.01
AGE	Continuous	0.047	0.002	22.11	< 0.01
ED3CAT × GENDER	<12 yrs × Male	0.004	0.102	0.04	0.970
	12 yrs × Male	0.201	0.087	2.20	0.026

Source: Analysis based on the 2006 HRS data.

[a] Reference categories for categorical predictors are GENDER (female); ED3CAT(>12 yrs).

B5.

a) The Error sum of squares is given by the formula $\sum(Y_{ij} - \widehat{Y_{ij}})^2 = SSE_R$, the Pure Error sum of squares $\sum(Y_{ij} - \overline{Y_j})^2$ =SSPE and the Lack of Fit sum of squares by $\sum(\overline{Y_j} - \widehat{Y_{ij}})^2 = \sum_j n_j(\overline{Y_j} - \widehat{Y_{ij}})^2$=SSLF

Show that SSE_R=SSLF+SSPE

b) A team of investigators at an Agricultural institute are investigating the effect of an amount of the particular chemical in a fertilizer on the amount of yield. They planted on 10 plots and collected the information in Table 2.

Table 2: Effect of amount of fertilizer on amount of yield output

Plot	1	2	3	4	5	6	7	8	9	10
% of chemical	20	10	20	50	50	10	30	40	30	40
yield	130	112	124	168	159	98	134	150	129	136

i) Construct an Analysis of Variance table and perform a lack of fit test on the data. Clearly show all calculations, hypotheses and conclusions in the answer.

ii) Calculate the coefficient of determination for the data and comment.

Compare your results with answer from b (i) **[10, 15, 5]**

B6

a) In a simple linear regression model, one can estimate the mean response at a value x_0 (a value within the range of X_i's) by using the model $\widehat{y(x_0)}=\widehat{\beta_0}+\widehat{\beta_1}(x_0)$.

Prove the following:

i) $E[\widehat{y(x_0)}]=\beta_0+\beta_1 x_0$ (is an unbiased estimator of the mean response);

ii) $Var[\widehat{y(x_0)}]=\frac{\sigma^2}{n}+(x_0-\overline{x})^2\frac{\sigma^2}{S_{xx}}$.

Hence state the 100(1-α)% confidence interval for the mean response.

b) An investigator is interested in the dependence of the speed of sound on temperature obtained the following results in Table 3.

Table 3: Dependence of speed of sound on temperature

X temperature [°C]	-20	0	20	50	100
Y Speed [m/s]	323	327	340	364	384

Obtain a 100(1-α)% confidence interval for the mean response, x_0=30 using the data above. **[(4, 8, 2), 16]**

END OF QUESTION PAPER