Strathmore
UNIVERSITY

Strathmore Institute of Mathematical Sciences

MASTER OF SCIENCE IN STATISTICAL SCIENCE

END OF SEMESTER EXAMINATION

STA 8103: Linear Models

Date: 20th December, 2021 **Time: 2 Hours**

Instructions

1. This examination consists of **FIVE** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

QUESTION ONE (20 MARKS)

(a) In order to analyze the effect of reducing nitrate loading in a Danish fjord, it was decided to formulate a linear model that describes the nitrate concentration in the fjord as a function of nitrate loading, it was further decided to correct for fresh water runoff. The resulting model was

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \; N(0, \sigma^2)$$

where Y_i is the natural logarithm of nitrate concentration, $x_{1,i}$ is the natural logarithm of nitrate loading, and $x_{2,i}$ is the natural logarithm of fresh water run off.

Comment on the following statements [using **TRUE** or **FALSE**] in the usual multiple linear regression model. Explain your comment.

(i) $\varepsilon_i = 0$ for all $i = 1, \cdots, n$ and β_j follows a normal distribution [2 Mark]

(ii) $\varepsilon_i = 0$ for all $i = 1, \cdots, n$ and x_j follows a normal distribution with $j = \{1, 2\}$. [2 Mark]

(b) If $\mathbf{a}$ is a $p \times 1$ vector of constants and y is a $p \times 1$ random vector with mean vector $\boldsymbol{\mu}$, show that [2 Marks]

$$E(\mathbf{a}'y) = \mathbf{a}'\boldsymbol{\mu}$$

(c) The following information is obtained from a sample data set.

$$n = 12, \quad \sum x = 66, \quad \sum y = 588, \quad \sum xy = 2104, \quad \sum x^2 = 350$$

Find the estimated regression line. [3 Marks]

(d) A commercial real estate company is interested in the relationship between properties' rental prices (Y), and the following predictors: building age, expenses/taxes, vacancy rates, and square footage. The results for a regression are given below.

Regression Statistics	
Multiple R	**0.7647**
R Square	**0.5847**
Standard Error	**1.1369**
Observations	**81**

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	**4**	**138.3269**	**34.5817**	**26.7555**	**0.0000**
Residual	**76**	**98.2306**	**1.2925**		
Total	**80**	**236.5575**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**12.2006**	**0.5780**	**21.1099**	**0.0000**	**11.0495**	**13.3517**
age	**-0.1420**	**0.0213**	**-6.6549**	**0.0000**	**-0.1845**	**-0.0995**
exp/tax	**0.2820**	**0.0632**	**4.4642**	**0.0000**	**0.1562**	**0.4078**
vacancy	**0.6193**	**1.0868**	**0.5699**	**0.5704**	**-1.5452**	**2.7839**
sqfoot	**0.0000**	**0.0000**	**5.7224**	**0.0000**	**0.0000**	**0.0000**

(i) Can the company conclude that rental rate is associated with any of these predictors? Give the test statistic and P- value for testing: [3 Marks]

H_0 : Average rental rate is not associated with any of the 4 predictors

H_A : Average rental rate is associated with at least one of the 4 predictors

(ii) What proportion of variation in prices is "explained" by the 4 predictors? [1 Marks]

(iii) Controlling for all other factors, we conclude age is
Positively / Negatively / Not associated with rental price. (Circle One) [1 Marks]

(e) Given the sample data set

x	5	2	-2	-2	3
y	6	0	2	-3	-1

Determine Covariance Matrix, $\sum$ [4 Marks]

(f) Explain how forward elimination method is used in variable selection [2 Marks]

QUESTION TWO (20 MARKS)

(a) State the three assumptions of ONE-WAY ANOVA [3 Mark]

(b) A researcher is interested in 3 independent variables; Murder, HS. Grad and Frost as predictors of Life.Exp. The output of analyzing the data is as follows.

```
Call:
lm(formula = Life.Exp ~ Murder + HS.Grad + Frost, data = statedata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5015 -0.5391  0.1014  0.5921  1.2268
Coefficients:
             Estimate Std. Error
(Intercept) 71.036379   0.983262
Murder      -0.283065   0.036731
HS.Grad      0.049949   0.015201
Frost       -0.006912   0.002447
---
```

(i) Test the hypothesis that "Murder" is not significant in predicting Life.Exp at α=0.05 . [5 Marks]

(ii) Find 95% confidence interval for the estimate for Murder [4 Marks]

(c) The least squares residuals for food expenditure of seven households are as presented in the table below.

x	y	$\hat{y} = 1.5050 + 0.2525x$	$e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
55	14	15.3925	-1.3925	1.9391
83	24	22.4625	1.5375	2.3639
38	13	11.1000	1.9000	3.6100
61	16	16.9075	-0.9075	0.8236
33	9	9.8375	-0.8375	0.7014
49	15	13.8775	1.1225	1.2600
67	17	18.4225	-1.4225	2.0235
$\sum x = 386$				$\sum e^2 = 12.7215$

Using the residuals for all n=7 observations, and the regression model ($y = b_1 + b_2 x + \varepsilon$), verify the following results:

(i) the estimated error variance, $\hat{\sigma}^2 = 2.5443$ [2 Marks]

(ii) the estimated variance of b_1, $\hat{var}(b_1) = 4.7273$ [2 Marks]

(iii) the estimated variance of b_2, $\hat{var}(b_2) = 0.001435$ [2 Marks]

(iii) the estimated covariance of b_1 and b_2, $\hat{cov}(b_1, b_2) = -0.0791$. [2 Marks]

QUESTION THREE (20 MARKS)

(a) The following ANOVA table, based on information obtained for four samples selected from four independent populations that are normally distributed with equal variances, has a few missing values. Find the missing values and complete the ANOVA table. [5 Marks]

Source of Variation	Degrees of freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between				
Within	15		9.2154	F=.....=4.07
Total	18			

(b) From time to time, unknown to its employees, the research department at Post Bank observes various employees for their work productivity. Recently this department wanted to check whether the four tellers at a branch of this bank serve, on average, the same number of customers per hour. The research manager observed each of the four tellers for a certain number of hours. The following table gives the number of customers served by the four tellers during each of the observed hours.

Teller A	Teller B	Teller C	Teller D
19	14	11	24
21	16	14	19
26	14	21	21
24	13	13	26
18	17	16	20
	13	18	

At the 5% significance level, test the null hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Assume that all the assumptions required to apply the one-way ANOVA procedure hold true. [15 Marks]

QUESTION FOUR (20 MARKS)

(a) The following information is obtained for a sample of 100 observations taken from a population.

$$SS_{xx} = 524.884, \quad s_e = 1.464, \quad \text{and} \quad \hat{y} = 5.48 + 2.50x$$

(i) Make a 98% confidence interval for B, the population parameter. [5 Marks]

(ii) Test at the 2.5% significance level whether B is positive. [6 Marks]

(b) A population data set produced the following information.

$$N = 460, \quad \sum x = 3920, \quad \sum y = 2650, \quad \sum xy = 26570,$$

$$\sum x^2 = 48530, \quad \text{and} \quad \sum y^2 = 39347$$

Find the linear correlation coefficient ρ. [5 Marks]

(c) Given the sample data set

x	1	3	5	3
y	1	1	3	3

Show that the covariance matrix, $\sum$, for the given data is

$$\sum = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

[4 Marks]

QUESTION FIVE (20 MARKS)

A diabetic is interested in determining how the amount of aerobic exercise impacts his blood sugar. When his blood sugar reaches 170 mg/dL, he goes out for a run at a pace of 10 minutes per mile. On different days, he runs different distances and measures his blood sugar after completing his run. Note: The preferred blood sugar level is in the range of 80 to 120 mg/dL. Levels that are too low or too high are extremely dangerous. The data generated are given in the following table.

Distance (miles)	2	2	2.5	2.5	3	3	3.5	3.5	4	4	4.5	4.5
Blood sugar (mg/dL)	136	146	131	125	120	116	104	95	85	94	83	75

(a) Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between distance run and blood sugar level?
. [3 Marks]

(b) Find the predictive regression equation of blood sugar level on the distance run. [6 Marks]

$$[\hat{y} = a + bx]$$

(c) Give a brief interpretation of the values a and b calculated in part (b) [4 Marks]

(d) Plot the predictive regression line on the scatter diagram of part (a) [4 Marks]

(e) Estimate the blood sugar level after a 10-mile run. Comment on this finding. [3 Marks]

You may Use the following formulae:

$\bar{x} = \frac{1}{n}\sum x$

$\sigma^2 = var(y) = E[y^2] - [E(y)]^2$

$\sigma_{ij} = cov(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)]$

$var(b_1) = \sigma^2\left[\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\right]; \qquad var(b_2) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}; \qquad cov(b_1, b_2) = \sigma^2\left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right]$

$$E[(\overrightarrow{y} - \overrightarrow{\mu})(\overrightarrow{y} - \overrightarrow{\mu})'] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} = \sum$$

$P_\rho = D_\sigma^{-1}\sum D_\sigma^{-1}$

$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}; \qquad SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}; \qquad SST = \sum y^2 - \frac{(\sum y)^2}{n}$

$b = \frac{SS_{xy}}{SS_{xx}}; \qquad s_e = \sqrt{\frac{SSE}{n-2}}$ (simple regression); $\qquad s_e = \sqrt{\frac{SSE}{n-k-1}}$ (multiple regression)

$r^2 = \frac{bSS_{xy}}{SS_{yy}}; \qquad t = \frac{b-B}{s_b}; \qquad t = \frac{\beta_i - \beta_i^*}{s_{\beta_i}}$ (multiple regression); $\qquad \gamma^2 = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}; \qquad t = r\sqrt{\frac{n-2}{1-r^2}}$

END OF PAPER