Strathmore
UNIVERSITY

INSTITUTE OF MATHEMATICAL SCIENCES
MSC (STATISTICAL SCIENCE)
END OF SEMESTER EXAMINATION
STA 8103: LINEAR MODELS
DATE: 17th December 2018 **TIME 2.5 Hours**

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

Question 1 (20 marks)

(a) In a multiple regression set up, show that $\hat{\beta} = X(X)^{-1}$ ***(5 Marks)***

(b) Consider a dataset with the dependent variable y and three predictor variables x1 x2.and x3. Two models were fit in order to check the usefulness of second-order terms in predicting y

Model 1:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \varepsilon$

Source of Variation	Sum of Squares	Degrees of freedom
Regression	1741.123	9
Error	46.21	9

Model 2:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

Source of Variation	Sum of Squares	Degrees of freedom
Regression	1707.158	3
Error	80.17	15

i. Using model 2 only, test the hypothesis that $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at α= 0.05 ***(3 Marks)***

ii. Based on the mallow's Cp statistic, would you say that model 2 is a good model? Explain your answer. ***(4 Marks)***

(c) Consider a model with a factor of 3 levels

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

Suppose we want to test the hypothesis that all the three populations are equal. Write down the hypothesis in terms of $H_0 : C\beta = 0$ hence indicate the elements in C and β ***(4 Marks)***

(d) Explain the process of conduction Analysis of covariance ***(4 Marks)***

Question 2 (20 marks)

(a) We model an individual's income at age 30 against the number of years of formal education (with a linear model). The following results were obtained: ***(4 marks)***

$n = 6, \quad \bar{x} = 4.67, \quad \sum (x - \bar{x})^2 = 29.33$

$s = 3.08, \quad \beta_0 = 3.04, \quad \beta_1 = 0.954$

Find 95% prediction interval for E(y) when x=7

(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 alpha=0.05 ***(5 marks)***

ii. Find 95% confidence interval for the estimate for Murder ***(3 marks)***

(c) Explain how backward elimination method is used in variable selection ***(4 marks)***

(d) List indicators of presence of multicollinearity ***(4 marks)***

Question 3 (20 marks)

(a) Consider an experiment with Factor A at two levels and Factor B at three levels. Assume that two observations are to be made for each combination of the levels of A and B. Assume also that both factors A and B are random. Let yijk be the kth observation corresponding to the ith level of A and the jth level of B.

(a) Write down the model for data yijk that includes interaction, and state all the assumptions. ***(6 marks)***

(b) Write down the hypothesis for testing factor A, factor B and interaction ***(6 marks)***

(b) Test the interaction effect given the results in the table below at alpha = 0.05 ***(4 marks)***

Source of variation	Degree of freedom	Sum of squares
A	1	205.4
B	2	2426.4
AXB (interaction)	2	108.3

Residual	54	712.1

Question 4 (20 marks)

(a) Show that the Poisson distribution belongs to the exponential family of distributions hence show that both the mean and variance are λ ***(14 marks)***

(b) A dataset consists of a dependent variable y and 3 independent variables x1, x2, x3. Based on the R output in appendix I, perform a stepwise selection to determine the best regression model for y. Give all the steps in detail. ***(6 marks)***

Appendix I

```
> problem = read.table("problem.dat", header=T);
> problem.lm1 = lm(y~x1, data=problem);
> summary(problem.lm1);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.4303     7.3788  16.863 1.92e-14 ***
x1            0.5554     0.6378   0.871    0.393
---

Residual standard error: 15.68 on 23 degrees of freedom
Multiple R-Squared: 0.03191,    Adjusted R-squared: -0.010
F-statistic: 0.7582 on 1 and 23 DF,  p-value: 0.3929

> problem.lm2 = lm(y~x2, data=problem);
> summary(problem.lm2);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  250.385     38.921   6.433 1.45e-06 ***
x2            -1.198      0.387  -3.094  0.00512 **
---

Residual standard error: 13.39 on 23 degrees of freedom
Multiple R-Squared: 0.2939,     Adjusted R-squared: 0.2632
F-statistic: 9.573 on 1 and 23 DF,  p-value: 0.005119

> problem.lm3 = lm(y~x3, data=problem);
> summary(problem.lm3);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 201.6314    16.5439  12.188 1.62e-11 ***
x3           -1.4093     0.3233  -4.359  0.00023 ***
---

Residual standard error: 11.79 on 23 degrees of freedom
Multiple R-Squared: 0.4524,     Adjusted R-squared: 0.4286
F-statistic:    19 on 1 and 23 DF,  p-value: 0.0002298

> problem.lm12 = lm(y~x1+x2, data=problem);
> summary(problem.lm12);

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 251.17351   44.82162   5.604 1.24e-05 ***
x1           -0.02266    0.59255  -0.038  0.96984
x2           -1.20301    0.42102  -2.857  0.00916 **
```

```
Residual standard error: 13.69 on 22 degrees of freedom
Multiple R-Squared: 0.2939,     Adjusted R-squared: 0.2298
F-statistic:  4.58 on 2 and 22 DF,  p-value: 0.02174

> problem.lm13 = lm(y~x1+x3, data=problem);
> summary(problem.lm13);

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 203.36968   19.99853  10.169 8.89e-10 ***
x1           -0.08366    0.51418  -0.163 0.872236
x3           -1.42629    0.34651  -4.116 0.000454 ***
---

Residual standard error: 12.05 on 22 degrees of freedom
Multiple R-Squared: 0.4531,     Adjusted R-squared: 0.4034
F-statistic: 9.113 on 2 and 22 DF,  p-value: 0.001309

> problem.lm23 = lm(y~x2+x3, data=problem);
> summary(problem.lm23);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.835     58.841   1.102  0.28243
x2             2.513      1.045   2.405  0.02503 *
x3            -3.685      0.991  -3.719  0.00120 **
---

Residual standard error: 10.73 on 22 degrees of freedom
Multiple R-Squared: 0.5664,     Adjusted R-squared: 0.527
F-statistic: 14.37 on 2 and 22 DF,  p-value: 0.0001018

> problem.lm123 = lm(y~x1+x2+x3, data=problem);
> summary(problem.lm123);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.2246    63.6825   0.930   0.3629
x1            0.1273     0.4762   0.267   0.7919
x2            2.5671     1.0869   2.362   0.0279 *
x3           -3.7085     1.0164  -3.649   0.0015 **
---

Residual standard error: 10.96 on 21 degrees of freedom
Multiple R-Squared: 0.5679,     Adjusted R-squared: 0.5062
F-statistic:   9.2 on 3 and 21 DF,  p-value: 0.0004392
```