Strathmore
UNIVERSITY

INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN STATISTICAL SCIENCES
END OF SEMESTER EXAMINATIONS
STA 8303-STATISICAL DATA MINING
DATE: September 7, 2022 Time: 3 Hours

Instructions
1. This examination consists of **FOUR** questions and an appendix to one of the questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1 (20 Marks**)**

a) Describe what a Markov Chain is and how Markov Chain Monte Carlo methods are used in statistical modeling.

(5 marks)

b) Suppose that a random variable Y has a Poisson(θ) distribution. That is,

$$f(y;\theta) = \frac{e^{-\theta}\,\theta^{y}}{y!}, y = 0,1, \ldots$$

Assuming a Gamma(α, β) prior-distribution for θ, derive expressions for the posterior density, posterior mean and variance for Y.
Hint: The density of a Gamma(α, β) variate *X* is:

$$f(x;\theta) = \frac{\Gamma(\alpha)}{\beta^{\alpha}} e^{-\beta x}\, x^{\alpha-1}, x > 0.$$

(7 marks)

c) Bagging, boosting, random forests and stochastic gradient boosting algorithms are ensemble methods that are often used to enhance the quality of a decision tree. Briefly describe each approach, clearly highlighting the enhancement that they give.

(8 marks)

Question 2 (20 Marks**)**
Consider the following data for a Bioassay:

	Blood clotting time (in seconds)	
Dose	Standard	Test (new)
0.025	67.5, 70.5, 67.5	68.8, 67.1, 67.0
0.050	62.1, 62.4	60.1, 59.4, 62.4

0.100	53.7, 51.6, 54.5	53.9, 51.7, 53.4, 52.7

a) Assuming a Normal likelihood, non-informative Gaussian priors for model parameters and a non-informative inverse-gamma prior distribution for the precision, the results of fitting concurrent, parallel and non-parallel lines to this data using are provided in Appendix 1. Write the WinBUGS code that would be used to obtain the results for the non-parallel lines model.

(8 marks)

b) Determine the best fitting model from the 3 models considered. Present a table indicating how the models of best fit was arrived at. [Include the residual mean sum of squares in this table]

(5 marks)

c) Interpret the estimated model coefficient for the best model considered and present posterior estimates of the parameters.

(5 marks)

d) Write the equations of the two non-parallel lines obtained.

(2 marks)

Question 3

a) Decision trees are procedures that are employed extensively in machine learning and statistical literature. Briefly explain how these procedures work and also mention aspects of their efficiency and reliability.

(7 marks)

b) A major problem associated with regression trees is instability. Explain what you understand by this problem.

(5 marks)

c) Describe the recursive partitioning algorithm and its utility in decision trees

(8 marks)

Question 4

a) Consider the following data:

x	1	2	4
y	1	3	1

Use quadratic spline interpolation to find the approximate value of y at x=3.

Hint:

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 4 & 0 & 2 & 0 & 1 & 0 \\ 0 & 4 & 0 & 2 & 0 & 1 \\ 0 & 16 & 0 & 4 & 0 & 1 \\ 4 & -4 & 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0.5 & -0.5 & -0.25 & 0.25 & 0.5 & -0.5 \\ -1.0 & 1.0 & 0.00 & 0.00 & 0.0 & -3.0 \\ -3.0 & 3.0 & 1.00 & -1.00 & -3.0 & 3.0 \\ 2.0 & -1.0 & 0.00 & 0.00 & 0.0 & 2.0 \\ 4.0 & -4.0 & 0.00 & 1.00 & 4.0 & -4.0 \end{pmatrix}$$

(10 Marks)

b) Consider the following data

$$(x_1, y_1) = (1,7), (x_2, y_2) = (2,23), (x_3, y_3) = (3,100)$$

Use polynomial interpolation to determine the value of the function at x=2.7.

Hint:

$$\begin{pmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 0.5 & -1 & 0.5 \\ -2.5 & 4 & -1.5 \\ 3.0 & -3 & 1.0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 0.5 & -1 & 0.5 \\ -2.5 & 4 & -1.5 \\ 3.0 & -3 & 1.0 \end{pmatrix}$$

(10 Marks)

APPENDIX 1

DIC-CONCURRENT LINES

Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
y	75.420	72.995	2.425	77.845
total	75.420	72.995	2.425	77.845

DIC-**PARALLEL LINES**

Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
y	75.220	71.710	3.510	78.729
total	75.220	71.710	3.510	78.729

DIC-NON-PARALLEL LINES

Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
y	78.675	74.699	3.976	82.651
total	78.675	74.699	3.976	82.651

POSTERIOR OUTPUT: CONCURRENT LINES

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	70.0	1.519	0.05005	65.79	70.27	72.09	1000	900
b[2]	-158.2	23.47	0.7186	-189.0	-163.7	-95.77	1000	900
s2	4.617	3.422	0.1034	1.515	3.68	14.65	1000	900

POSTERIOR OUTPUT: PARALLEL LINES

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	70.56	1.562	0.05009	66.41	70.88	72.79	1000	900
b[2]	-158.0	23.4	0.7875	-189.9	-163.0	-98.72	1000	900
b[3]	-1.038	1.023	0.03237	-3.051	-1.024	0.9043	1000	900
s2	4.636	4.03	0.1477	1.428	3.508	15.35	1000	900

POSTERIOR OUTPUT: NON-PARALLEL LINES

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	69.45	1.806	0.06232	65.07	69.66	72.32	1000	900
b[2]	-139.3	26.33	0.9038	-182.5	-142.8	-77.65	1000	900
b[3]	0.952	1.65	0.04921	-2.024	0.8281	4.546	1000	900
b[4]	-31.76	20.3	0.6435	-76.28	-30.47	5.48	1000	900
s2	5.639	4.385	0.1568	1.708	4.232	17.63	1000	900

APPENDIX 2

LOGISTIC REGRESSION MODEL

DIC
Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
Y	61.536	59.489	2.047	63.583
total	61.536	59.489	2.047	63.583

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	-4.605	0.1513	0.01	-4.905	-4.601	-4.312	1000	900
b[2]	0.05341	0.008706	5.687E-4	0.03617	0.05324	0.07092	1000	900

model is syntactically correct
data loaded
model compiled
model is initialized

RANDOM EFFECTS MODEL

DIC
Dbar = post.mean of -2logL; Dhat = -2LogL at post.mean of stochastic nodes

	Dbar	Dhat	pD	DIC
Y	58.512	53.639	4.874	63.386
total	58.512	53.639	4.874	63.386

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
b[1]	-4.675	0.2324	0.02112	-5.126	-4.672	-4.249	1000	900
b[2]	0.05777	0.01278	0.001145	0.03399	0.05776	0.08321	1000	900
sigma	0.1025	0.0666	0.002992	0.02645	0.08672	0.2655	1000	900
u[1]	0.01318	0.08738	0.004514	-0.1497	0.0111	0.2045	1000	900
u[2]	0.0452	0.09858	0.006998	-0.1083	0.03301	0.2656	1000	900
u[3]	0.00462	0.07617	0.002334	-0.1469	0.002108	0.1654	1000	900
u[4]	-0.02898	0.09073	0.002719	-0.2352	-0.01517	0.1235	1000	900
u[5]	0.006685	0.0916	0.004234	-0.1908	0.006101	0.2016	1000	900
u[6]	-0.09702	0.1007	0.003923	-0.3432	-0.07819	0.04866	1000	900
u[7]	0.07773	0.115	0.006063	-0.0905	0.05434	0.3431	1000	900
u[8]	-0.03649	0.08688	0.004385	-0.2316	-0.02859	0.1136	1000	900