National University of Computer and Emerging Sciences

FAST School of Computing **Fall-2023** **Islamabad Campus**

MT-2002: Statistical Modeling

Serial No:

2nd Sessional Exam

Solution

Total Time: 1 Hour

Total Marks: 45

Saturday, 4th November, 2023

Course Instructors

Dr. Noreen Jamil,
Dr. Shahnawaz Qureshi,
Dr. Imran Ashraf,
Muhammad Almas Khan

____
Signature of Invigilator

Student Name **Roll No.** **Course Section** **Student Signature**

DO NOT OPEN THE QUESTION BOOK OR START UNTIL INSTRUCTED.

Instructions:

1. Attempt on question paper. Attempt all of them. Read the question carefully, understand the question, and then attempt it.
2. No additional sheet will be provided for rough work. Use the back of the last page for rough work.
3. If you need more space, write on the back side of the paper and clearly mark the question and part number etc.
4. After asked to commence the exam, please verify that you have **Seven (7)** different printed pages including this title page. There are total of **5** questions.
5. Calculator sharing is strictly prohibited.
6. Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.

	Q-1	Q-2	Q-3	Q-4	Q-5	Total
Marks Obtained						
Total Marks	12	6	8	12	7	45

Question 1: (a) Dr. Noreen
Question 2 & 3: Dr. Imran

Question 1 [8+4=12 Marks] Dr. Noreen

a) A sample of students are measured for height each year for 3 years. After the third year, you want to fit a linear regression predicting height using year as a predictor. Write down the Bayesian model definition (**using statistical notations**) for this regression, using any variable names and priors you choose.

Since height is continuous variable, and we only have one independent variable (predictor), we can use

robust linear regression model. Since we have continuous data, it is better to go for robust linear regression model as it deals with outliers as well.

Priors:-

$$\alpha \sim \text{Normal}(\mu_\alpha, \sigma_\alpha)$$
$$\beta \sim \text{Normal}(\mu_\beta, \sigma_\beta)$$
$$\epsilon \sim \text{Half Cauchy}(\sigma_\epsilon)$$
$$v \sim \text{Exponential}(\lambda)$$
$$\mu = \alpha + \beta x$$

Likelihood:-

$$y \sim \text{Student T}(mu = \mu, sd = \epsilon, v)$$

The choice of our priors is simple as we have continuous data. "ϵ" can be calculated using half normal or half cauchy but as we have to deal with outliers so we have to use half cauchy here. "v" is the normality parameter.

b) Assume we want to use the above model for another dataset in which x and y have a polynomial relationship of degree 3. How can we modify the above linear model. Rewrite it in statistical notation to accommodate this polynomial relationship.

Ans: modified version of both of the models are here!

Simple linear model

$\alpha \sim$ Normal $(\mu_\alpha, \sigma_\alpha)$
$\beta_1 \sim$ Normal $(\mu_\beta, \sigma_\beta)$ (1 mark)
$\beta_2 \sim$ Normal $(\mu_\beta, \sigma_\beta)$ (1 mark)
$\beta_3 \sim$ Normal $(\mu_\beta, \sigma_\beta)$ (1 mark)
$\varepsilon \sim$ Half Cauchy $(\mu_\varepsilon\ \varepsilon_\varepsilon)$
$\mu = \alpha + \beta_1 (x) + \beta_2 (x)^2 + \beta_3 (x)^3$ (1 mark)
$y \sim$ Normal (mu= μ, sd= ε)

Robust linear model

$\alpha \sim$ Normal $(\mu_\alpha, \sigma_\alpha)$
$\beta_1 \sim$ Normal $(\mu_\beta, \sigma_\beta)$ (1 mark)
$\beta_2 \sim$ Normal $(\mu_\beta, \sigma_\beta)$ (1 mark)
$\beta_3 \sim$ Normal $(\mu_\beta, \sigma_\beta)$ (1 mark)
$\varepsilon \sim$ Half Cauchy $(\mu_\varepsilon\ \varepsilon_\varepsilon)$
$\mu = \alpha + \beta_1 (x) + \beta_2 (x)^2 + \beta_3 (x)^3$ (1 mark)
v= Exponential(λ)
$y \sim$ Student t (mu= μ, sd= ε, v)

Question 2 [6 Marks] Dr. Imran Ashraf

For the logit() and logistic() function, provide:

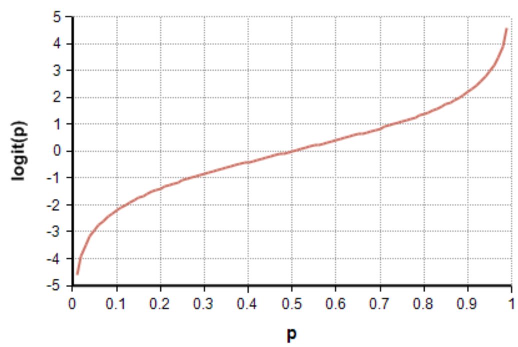
a. [1x2 marks] mathematical expression for both

$$logit(p) = log\frac{p}{1-p}$$

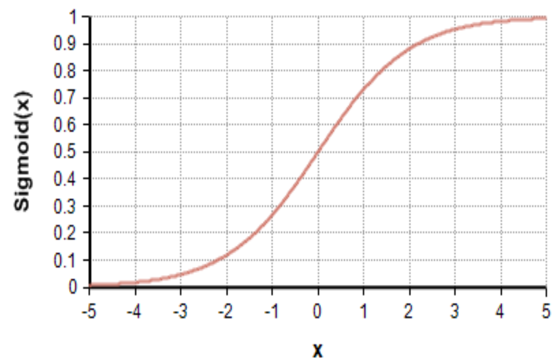
$$logistic(x) = sigmoid(x) = \frac{1}{1+exp(-x)}$$

b. [2x2 marks] properly labeled plot for both

Logit plot



logistic/sigmoid plot



Question 3 [8 Marks] Dr. Imran Ashraf

a) [2 marks] What will be the expression for the mean (μ) of a regression model which uses two features (x1 and x2) and also involves an interaction term.

The expression for the mean (μ) having two features is given below.

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

b) [2 marks] If we have to use this model for logistic regression, how can we compute probability by using this μ?

We can use the logistic function for the calculation of probability by:

$$logistic(\mu) = logistic(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\beta 3x1x2)}}$$

c) [4 marks] Derive the relationship for the decision boundary of this model to be used for logistic regression.

$$logistic(\mu) = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\beta 3x1x2)}}$$

For the decision boundary, Logistic(μ) should be 0.5, which gives.

$$0.5 = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\beta 3x1x2)}}$$

Simplifying this gives the desired decision boundary to be:

$$x2 = -\frac{\beta 0 + \beta 1 x1}{\beta 2 + \beta 3 x1}$$

Question 4 [8+4=12 Marks] M. Almas

Imagine you are a data analyst at a software development company. Applying Bayesian models for classification task. Consider the following dataset features and classes.

Project ID	Coding Efficiency (CE)	Error Rate (ER)	Project Management Score (PMS)	Project Outcome
1	80	0.1	90	High-Quality
2	75	0.2	85	High-Quality
3	60	0.9	60	Average-Quality
4	70	0.35	88	Low-Quality

You initially attempted binary classification i.e., High-Quality & Low-Quality using the following simple logistic Bayesian model, based on one feature Error Rate: let x = **Error Rate**

Model for Binary Classification based on one feature:

$\alpha \sim$ Normal (μ, σ)
$\beta \sim$ Normal (μ, σ)
$\mu = \alpha + \beta\,(x)$
$\theta = \text{sigmoid}(\mu)$
$bd = -\alpha/\beta$
$y \sim \text{Bern}(\theta)$

However, as you explore the dataset, you realize that trying to predict software project outcomes using just one feature is insufficient.

a) Modify above given simple logistic model and convert it into multiple logistic models for binary classification using **Error Rate**, **Coding efficiency as** your features from above dataset. Specify the changes & rewrite the modified version model as mentioned in above notation.

Ans: Suppose, $\mathbf{x_1}$ = Error Rate and $\mathbf{x_2}$ = Coding efficiency then the modified version of model is given as below along with the ,

$\alpha \sim$ Normal (μ, σ)
$\beta_1 \sim$ Normal (μ, σ) (1 marks)
$\beta_2 \sim$ Normal (μ, σ) (1 marks)
$\mu = \alpha + \beta_1\,(x_1) + \beta_2\,(x_2)$ (3 marks)
$\theta = \text{sigmoid}(\mu)$
$bd = -\frac{\alpha}{\beta_0} + \left(-\frac{\beta_0}{\beta_1} x_1\right)$ (3 marks)
$y \sim \text{Bern}(\theta)$

b) Continuing from the multiple logistic regression model in (a), Rewrite the modified version model by converting it to Bayesian model for multi-class classification. Highlight the modifications i.e., **function** and **likelihood** required to accommodate all possible project outcome categories in the dependent variable.

Ans: The modified model as per the description above is given as below.

$\alpha \sim$ Normal (μ, σ)

$\beta 1 \sim$ Normal (μ, σ)
$\beta 2 \sim$ Normal (μ, σ)
$\mu = \alpha + \beta 1\,(x1) + \beta 2\,(x2)$
θ =SoftMax(μ) 2 marks
y ~ Categorical (θ) 2 marks

Question 5 [1*7=7 Marks]

Assume you implemented a multiple logistic Bayesian regression model. Where x1, x2 are the independent variables. Consider the following arviz summary of the Multiple logistic Bayesian model. Interpret the following results. Select correct option for MCQs in context of multiple logistic Bayesian model (from 1-4 MCQs).

	mean	**sd**
α	-9.12	4.61
β[0]	4.65	0.87
β[1]	-5.16	0.95

1. In logistic regression, what does the coefficient β[0] represent?
✓β[0] encodes the increase in log-odds units by a unit increase of the predictor variable x1.
☐β[0] encodes the increase in log-odds units by a unit increase of the predictor variable x2.
☐β[1] encodes the increase in log-odds units by a unit decrease of the predictor variable x1.
☐β[0] is the standard error of the estimate.

2. Look at β[0] in the above table of summary, what does this coefficient tell us about the relationship between the predictor variable and the outcome?
☐A unit increase in the predictor variable x1 reduces the odds of the outcome by -4.36 units.
✓A unit increase in the predictor variable x1 reduces the log odds of the outcome by 4.65 units.
☐A unit increase in the predictor variable increases the odds of the outcome by -4.26 units.
☐A unit increase in the predictor variable x2 reduces the odds of the outcome by -4.16 units.

3. Look at β[1] in above table of summary, what does this coefficient tell us about the relationship between the predictor variable and the outcome?
☐A unit increase in the predictor variable x1 reduces the odds of the outcome by -5.36 units.
☐A unit increase in the predictor variable x2 reduces the odds of the outcome by -5.16 units.
☐A unit increase in the predictor variable x1 increases the odds of the outcome by -5.26 units.
✓A unit increase in the predictor variable reduces the log odds of the outcome by -5.16 units.

4. In a linear regression equation y=mx+b, what does "m" represent?
✓Slope of the regression line
☐Intercept of the regression line
☐Variance of the data points
☐Mean of the data points

5. In a polynomial regression equation of degree 2 (quadratic), what does the equation look like?

☐y = a + bx

✓y = a + bx + cx^2

☐y = a + bx + cx^3

☐y = a + bx + cx^0

6. What is the primary advantage of using a higher degree polynomial in polynomial regression?

☐It reduces over-fitting.

☐It makes the model simpler.

✓It allows the model to capture more complex relationships.

☐It speeds up the training process.

7. Suppose the Pearson correlation between V1 and V2 is zero. In such a case, is it right to conclude that V1 and V2 do not have any linear relation between them?

✓TRUE

☐FALSE

Complete marks distribution

	Questions	parts	Marks of each part
	1	a	8
		b	4
	2	a	2
		b	4
	3	a	2
		b	2
		c	4
	4	a	8
		b	4
	5	7 MCQs	7
Total	5 questions	9 parts and 7 mcqs	45