National University of Computer and Emerging Sciences

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

MT-2002: Statistical Modeling

Serial No:

Final Exam
Part I (1 hour)
Part II (2 hours)
Total Time: 3 Hours
Total Marks: 92

Thursday, 21st December, 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. Read the question carefully, understand the question, and then attempt it.
2. No additional sheet will be provided for rough work.
3. Verify that you have **Eighteen (18)** different printed pages including this title page & answer sheet for MCQs. There are **Five (5)** questions.
4. **Part I** is **40 MCQs.** You need to answer MCQs part using answer sheet provided to you in the last of **part I.**
5. **Part II** Consists of Subjective Question.
6. Calculator sharing is strictly prohibited.
7. Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.
8. Ensure that you do not have any electronic gadget (like mobile phone, smart watch, etc.) with you.

	Q-1 (40 MCQs)	Q-2	Q-3	Q-4	Q-5	Total
Marks Obtained						
Total Marks	40	8	20	3	21	92

Part I [40 marks]

Attention!
Answer 40 MCQs, use answer sheet on last page of this part & fill clearly the circle with a pen/marker.

Question 1 [40 Multiple Choice Questions]

1. The Pearson correlation coefficient measures the:

   a. Strength of a linear relationship between two variables.
   b. Magnitude of the difference between two groups.
   c. Likelihood of an event occurring.
   d. Proportion of explained variance in a regression model.

2. What does the determination coefficient (R-squared) represent in the context of linear regression?

   a. The correlation between independent and dependent variables.
   b. The proportion of the total variation in the dependent variable explained by the model.
   c. The ratio of the residuals to the predicted values.
   d. The standard error of the regression coefficients.

3. In multiple linear regression using PyMC, what is the primary difference compared to simple linear regression?

   a. Multiple linear regression can only handle two variables.
   b. Multiple linear regression involves one dependent variable.
   c. Multiple linear regression models the relationship between multiple independent variables and one dependent variable.
   d. Multiple linear regression cannot handle categorical variables.

4. What is the primary concern associated with confounding variables in multiple linear regression?

   a. Confounding variables introduce bias and can lead to incorrect conclusions about the relationships between variables.
   b. Confounding variables always improve the accuracy of the regression model.
   c. Confounding variables are irrelevant and have no impact on the regression results.
   d. Confounding variables are only applicable in simple linear regression, not in multiple linear regression.

5. How does multi-collinearity impact the interpretation of regression coefficients in multiple linear regression?

   a. Multi-collinearity has no effect on the interpretation of regression coefficients.
   b. It makes the regression coefficients more reliable and stable.
   c. multi-collinearity inflates standard errors, making it challenging to discern the individual impact of each variable.

d. Multi-collinearity is not an issue.

6. What is the primary purpose of using polynomial regression in statistical modeling?

   a. To model linear relationships between variables
   b. To model non-linear relationships between variables
   c. To simplify data visualization
   d. To reduce the complexity of the model

7. In polynomial regression, what does the degree of the polynomial represent?

   a. The number of variables in the model
   b. The size of the dataset
   c. The order of the polynomial equation
   d. Nothing

8. When fitting a polynomial regression model, what is the potential risk of choosing a higher degree polynomial?

   a. Nothing happened.
   b. Overfitting
   c. Increased interpretability
   d. Reduced computational complexity.

9. Which of the following functions is commonly used in logistic regression to model binary outcomes?

   a. Parabolic function
   b. Cos function
   c. Sigmoid function
   d. Sine function

10. In logistic regression, the logistic function transforms the linear combination of input features and weights into

    a. A straight line
    b. A step functions.
    c. A sigmoid curve
    d. An exponential curve

11. What type of dependent variable is most suitable for logistic regression?

    a. Continuous
    b. Binary
    c. Categorical with multiple levels
    d. No variable needed.

12. "log odds" in logistic regression are?

    a. The logarithm of the dependent variable.
    b. The logarithm of the odds ratio.
    c. Nothing
    d. The derivative of regression coefficients.

13. In logistic regression, the odds ratio represents:

    a. The change in the mean of the dependent variable for a one-unit change in the independent variable.
    b. The change in the log-odds of the dependent variable for a one-unit change in the independent variable.
    c. The probability of success is in the dependent variable.
    d. The intercept of the regression line.

14. What challenge may arise in multiple logistic regression when dealing with a large number of predictor variables?

    a. Improved model interpretability.
    b. Enhanced model stability.
    c. The class imbalance.
    d. Increased convergence speed

15. What is a common issue associated with unbalanced classes in logistic regression?

    a. It has no impact on model performance.
    b. It can lead to biased model predictions.
    c. It results in faster model convergence.
    d. Boost accuracy.

16. In statistical modeling, the SoftMax function is commonly used for

    a. Predicting numeric value
    b. Multinomial (multiclass) classification problems.
    c. Regression analysis.
    d. Time series forecasting

17. If you input a vector of five real values into a SoftMax function, the output will be:

    a. The same as the input vector.
    b. The square root of the input vector.
    c. A normalized probability distribution over the five values.
    d. The sum of the input vector values.

18. What does the SoftMax function do when applied to a vector of real numbers?

    a. Squashes the values to a range between 0 and 1.
    b. Computes the exponential of each element and normalizes the result.
    c. Applies a linear transformation to the vector.
    d. Converts the vector into a binary output.

19. In statistical modeling, when is the SoftMax function preferred over the sigmoid function?

    a. When dealing with binary classification problems.
    b. When there are multiple classes and each observation can belong to only one class.
    c. When the relationship between input features and output is linear.
    d. When modeling a probability distribution for a single binary outcome.

20. In statistical modeling, what is a key distinction between the sigmoid and softmax functions?

a. Sigmoid is used for binary classification, while softmax is used for multiclass classification.
b. Sigmoid is faster in terms of computation.
c. Softmax is only applicable to linear regression.
d. Sigmoid is more robust to outliers.

21. Poisson regression is most appropriate for modeling dependent variable:

a. Continuous outcomes.
b. Binary outcomes.
c. Count data.
d. Time series data.

22. In Poisson regression, the dependent variable is assumed to follow which type of distribution?

a. Normal distribution.
b. Poisson distribution.
c. Binomial distribution.
d. Exponential distribution.

23. When might a Poisson regression model be more appropriate than logistic regression for modeling count data?

a. When the outcome variable is binary.
b. When the outcome variable is categorical with multiple levels.
c. When the outcome variable represents the number of occurrences within a fixed unit of time or space.
d. When the relationship between the input variables and the outcome is linear.

24. In a statistical modeling context, the Zero-Inflated Poisson (ZIP) model is primarily designed for situations where.

a. The dependent variable follows a normal distribution.
b. There is excessive variability in the dependent variable.
c. There is an excess of zero counts in the dependent variable.
d. The relationship between variables is linear.

25. In posterior predictive checks (PPC), we are comparing p values for two models, we should expect p -value around?
a. 0.5
b. 200
c. 150
d. 100

26. What happens when a statistical model has too many parameters?

a. The model becomes more accurate.
b. The model overfits the training data.
c. The model underfits the training data.
d. The model requires less computational resources.

27. What happens when a statistical model has too few parameters?

a. The model becomes more accurate.
b. The model overfits the training data.

c. The model underfits the training data.
d. The model requires less computational resources.

28. When comparing models using Arviz & IC = WAIC/Loo, with “**deviance**” as scale, then

a. Small value represents a better model.
b. Infinite value will represent a better model.
c. Only positive value will represent a better model.
d. Every value represents a better model.

29. When comparing models using Arviz & IC = WAIC/Loo, with “**negative log**” as scale, then

a. Large value represents a better model.
b. Small value represents a better model.
c. Only positive value will represent a better model.
d. Every value represents a better model.

30. When comparing models using Arviz via IC = WAIC/Loo, with “**log**” as scale, then

a. Large value represents a better model.
b. Small value represents a better model.
c. Only positive value will represent a better model.
d. Every value represents a better model.

31. people often compute bayes factors as the ratio of two marginal likelihoods i.e. BF = $p(y|M_0 \frac{}{p(y|M_1)}$

a. When BF > 1, M0 explains data better than M1.
b. When BF <1, M0 explains data better than M1.
c. When BF =1, M0 explains data better than M1.
d. When BF =-1, M0 explains data better than M1.

32. Let suppose we are comparing two models i.e. M0 & M1 via Bayes Factor (BF), then which of the following is correct?

a. If BF = 0, moderate evidence favoring M0 than M1
b. If BF =9, moderate evidence favoring M0 than M1
c. If BF =9 strong evidence favoring M0 than M1
d. If BF =0 very strong evidence favoring M0 over M1

33. If Bayes Factor is > 100

a. extreme evidence favoring M0 than M1
b. very strong moderate evidence favoring M0 than M1
c. moderate strong evidence favoring M0 than M1
d. anecdotal very strong evidence favoring M0 over M1

34. What does entropy measure in statistical modeling?

a. The average difference between observations.
b. The spread of data points.
c. The uncertainty or disorder in a distribution.
d. The correlation between variables.

35. Kullback-Leibler divergence measures:

    a. The similarity between two probability distributions.
    b. The sum of squared differences between data points.
    c. The average absolute difference between observations.
    d. The correlation coefficient between two variables.

36. What does Kullback-Leibler divergence do that entropy doesn't?

    a. Measures disorder in a single distribution.
    b. Compares two distributions for similarity.
    c. Quantifies the spread of data points.
    d. Assesses the average difference between observations.

37. What is the primary purpose of model averaging in statistical modeling?

    a. To make model simpler.
    b. To combine predictions from multiple models.
    c. To overfit a single model.
    d. To increase the training time of models

38. What is the primary goal of A/B testing in the context of website conversion rates?

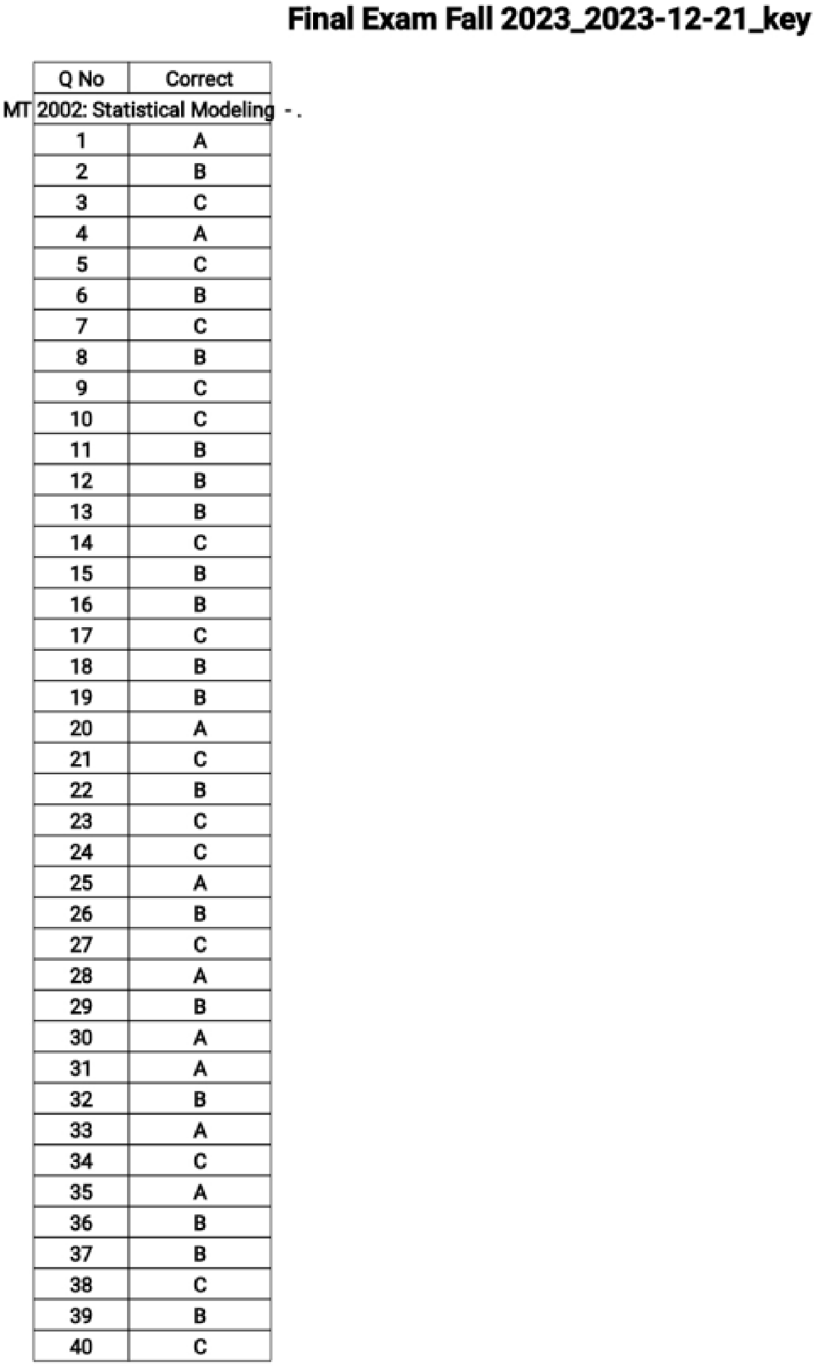
    a. To increase website traffic.
    b. To identify areas for website redesign.
    c. To compare two or more versions of a webpage and determine which performs better in terms of conversion.
    d. To analyze the total revenue generated by the website.

39. If Version A of a webpage has a conversion rate of 8% and Version B has a conversion rate of 12%, what can be concluded?

    a. Version A is better.
    b. Version B is better.
    c. Both versions are equally effective.
    d. The test results are inconclusive.

40. What role does a control group play in A/B testing?

    a. To introduce bias into the experiment.
    b. To ensure that all visitors see the same version of the webpage.
    c. To serve as a baseline for comparison and measure the effect of changes on the conversion rate.
    d. To eliminate the need for statistical analysis.



Final Exam Fall 2023_2023-12-21_key

Q No	Correct
MT 2002: Statistical Modeling - .	
1	A
2	B
3	C
4	A
5	C
6	B
7	C
8	B
9	C
10	C
11	B
12	B
13	B
14	C
15	B
16	B
17	C
18	B
19	B
20	A
21	C
22	B
23	C
24	C
25	A
26	B
27	C
28	A
29	B
30	A
31	A
32	B
33	A
34	C
35	A
36	B
37	B
38	C
39	B
40	C

Part II [52 marks]

Question 2 [3+3+2= 8 marks]

Recall the concept discussed in the class about regression problems, particularly when transitioning from one independent variable to multiple independent variables. Now, let's analyze the two plots provided, focusing on the challenges that arise when moving from one independent variable to more than one independent variable. Respond to the questions posed for each plot in the following sections.

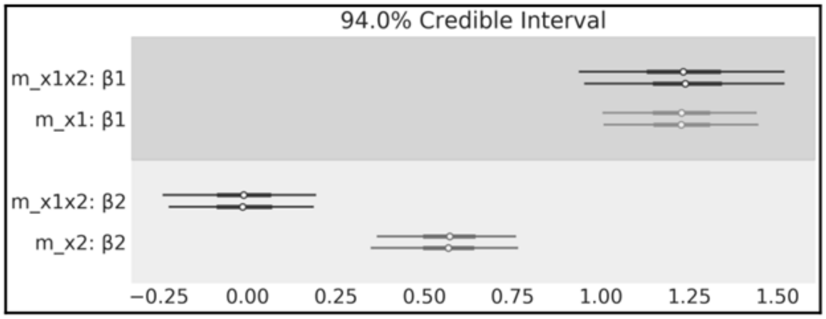


Figure 3.22

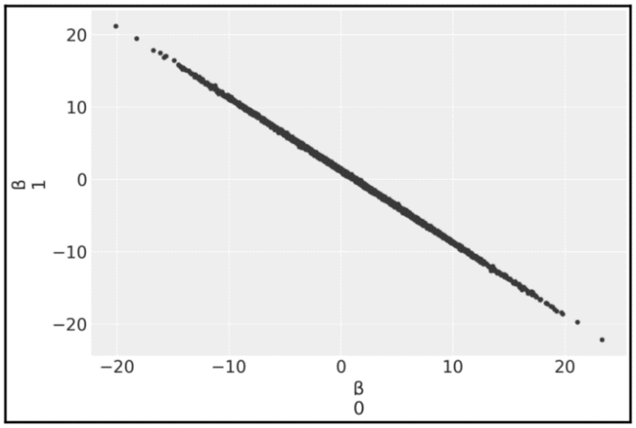
a) [3 marks] What is the term used to describe the challenge encountered as shown above when dealing with more than one independent variable?

When there are more than one independent variables, we face some challenges, in the above plot, we can observe one of the challenges: a confounding variable related to the multiple linear regression model.

b) [3 marks] In the depicted plot above showcasing the posterior distribution of β1 and β2 for variables X1 and X2, estimated by different models. Which variable is confounding? X1 or X2.

Variable X1 is confounding.

c) [2 marks] Consider the following plot.



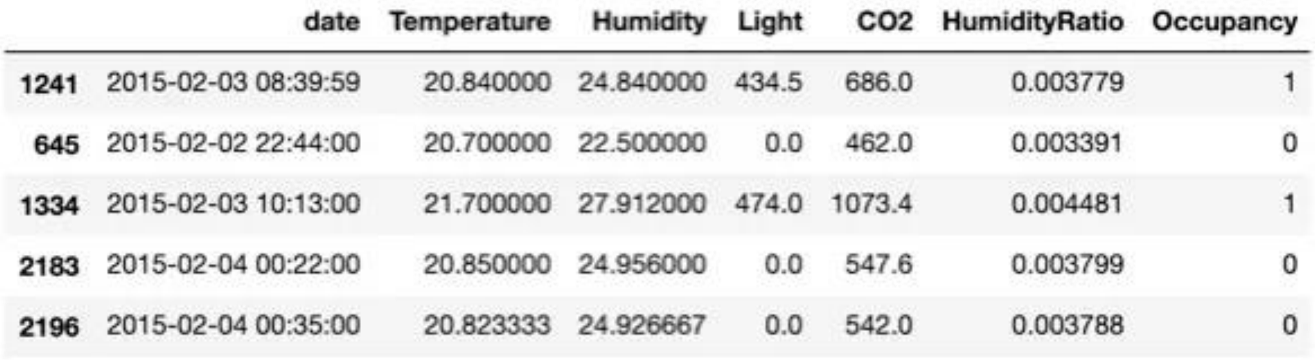
When dealing with more than two variables in a regression model, what is the specific challenge indicated in the above plot?

When dealing with multiple independent variables in a regression model, we may encounter various challenges, and one of them is multicollinearity. In the above plot, the coefficients (beta) of two variables are correlated, leading to the model experiencing multicollinearity.

Question 3 [4+4+4+2+4+2=20 marks]

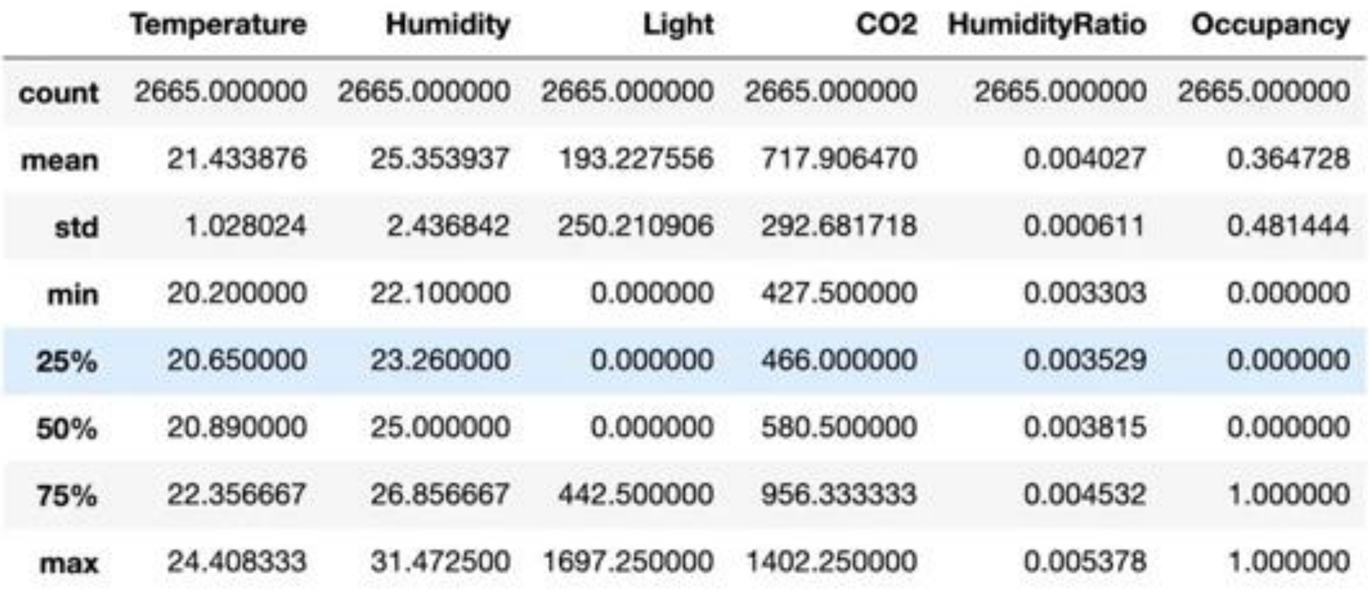
In this question you will build a **binary classification model** to predict whether a room is occupied or not (0 for unoccupied and 1 for occupied). The dataset contains several features such as light, temperature, humidity, and CO2 levels. The goal is to detect a room's occupancy from these variables.

Here is a small part of this dataset.



	date	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
1241	2015-02-03 08:39:59	20.840000	24.840000	434.5	686.0	0.003779	1
645	2015-02-02 22:44:00	20.700000	22.500000	0.0	462.0	0.003391	0
1334	2015-02-03 10:13:00	21.700000	27.912000	474.0	1073.4	0.004481	1
2183	2015-02-04 00:22:00	20.850000	24.956000	0.0	547.6	0.003799	0
2196	2015-02-04 00:35:00	20.823333	24.926667	0.0	542.0	0.003788	0

Figure below shows the output of df.describe()



	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
count	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000
mean	21.433876	25.353937	193.227556	717.906470	0.004027	0.364728
std	1.028024	2.436842	250.210906	292.681718	0.000611	0.481444
min	20.200000	22.100000	0.000000	427.500000	0.003303	0.000000
25%	20.650000	23.260000	0.000000	466.000000	0.003529	0.000000
50%	20.890000	25.000000	0.000000	580.500000	0.003815	0.000000
75%	22.356667	26.856667	442.500000	956.333333	0.004532	1.000000
max	24.408333	31.472500	1697.250000	1402.250000	0.005378	1.000000

a). [4 marks] In your first model, you use two features (Humidity and CO2). Write your binary classification model using statistical notation. Use appropriate priors.

ANSWER

Let suppose. x_1 = Humidity
x_2 = CO_2
y = Occupancy
α = Intercept
β_1= Coefficient for Humidity
β_2= Coefficient for CO_2

$\alpha \sim$ Normal $(\mu_\alpha, \sigma_\alpha)$ (0.5 marks)
$\beta_1 \sim$ Normal $(\mu_{\beta1}, \sigma_{\beta1})$ (0.5 marks)
$\beta_2 \sim$ Normal $(\mu_{\beta2}, \sigma_{\beta2})$ (0.5 marks)
$\mu =$ $\alpha + \beta_1 (x_1) + \beta_2(x_2)$ (0.5 marks)
$\theta =$ sigmoid(μ) (0.5 marks)
bd = $-\frac{\alpha}{\beta_1} +(- \frac{\beta_1}{\beta_2} x_1)$ (1 mark)
$y \sim$ Bern(θ) (0.5 marks)

b) [2+2] Use suitable values for parameters instead of using generic symbol names in prior distribution. Provide reasons for your choice.

ANSWER

By using the data given in df.describe() which shows a summary of mean and std for each feature, we can use informed value for μ, σ as shown below:

$\alpha \sim$ Normal (0.36, 0.48) (2 marks)
$\beta_1 \sim$ Normal (25.35, 2.43) (1 marks)
$\beta_2 \sim$ Normal (717.90, 292.68) (1 marks)

c) [4] Derive the expression for boundary decision.

ANSWER

$$\text{logistic}(\mu) = \frac{1}{1+e^{-(\beta 0+\beta 1 x 1+\beta 2 x 2)}}$$

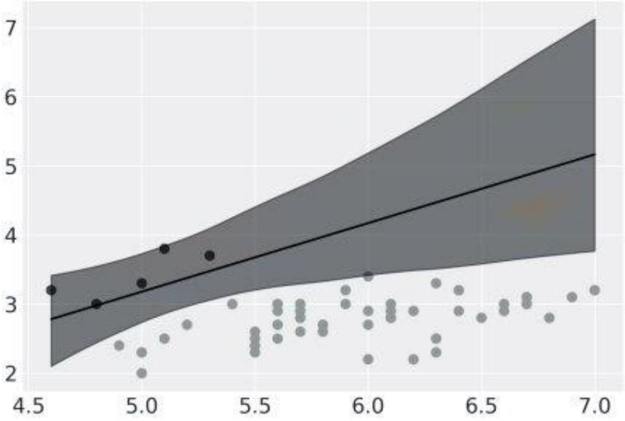
For the decision boundary, Logistic(μ) should be 0.5, which gives.

$$0.5= \frac{1}{1+e^{-(\beta 0+\beta 1 x 1+\beta 2 x 2)}}$$

Simplifying this gives the desired decision boundary to be:

$$x_2 = \frac{-\beta 1}{\beta 2} + \left(\frac{-\beta 1}{\beta 2} x_2\right)$$

You use two features and based on that you get the following plot showing the decision boundary.

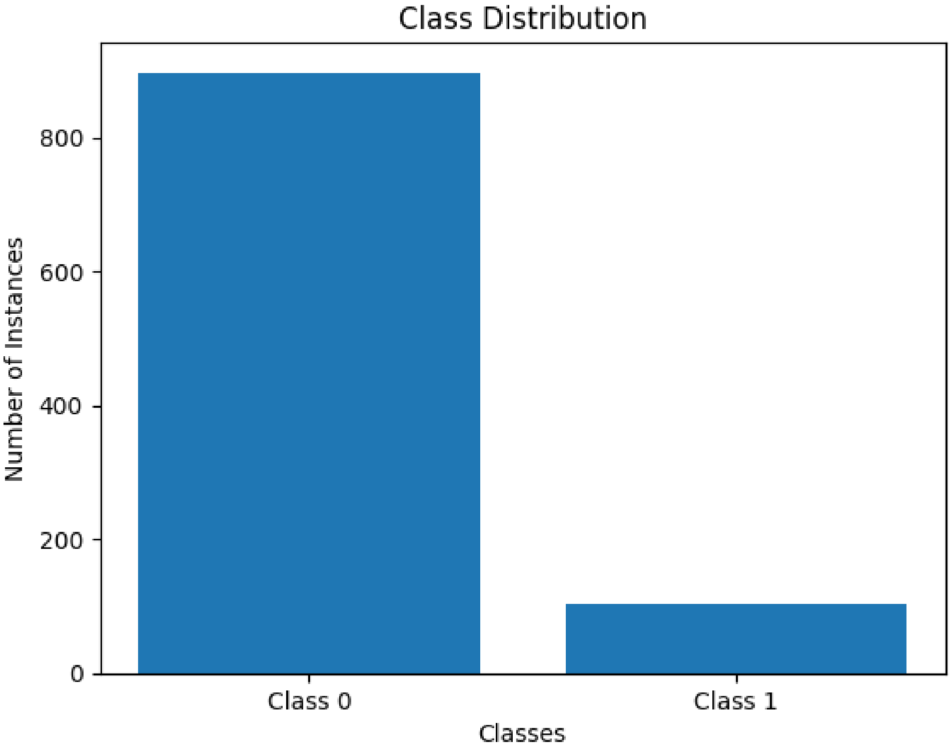


d) [2 marks]: As you can see, the boundary decision is shifted towards one class. What might be the reason for this?

The challenge, known as "class imbalance," arises when unequal class representation biases predictions, posing a risk of biased results for the majority class and compromising performance results of the minority classes.

e) [4 marks]: Which single plot you can generate to validate your answer in the previous part? Provide a properly labeled example plot for this which is representative of this situation.

A simple bar graph should be generated to show the class distribution as shown by the properly labeled graph below which clearly depicts the class imbalance.



f) [1+1 marks]: Mention two solutions (names only) which can be utilized to solve this problem.

[1 mark] **Dataset Balancing**: Balance the classes by getting more samples or synthetically generating more data (dataset augmentation).

[1 mark] **Prior Information Inclusion**: use informative priors.

Question 4 [3 marks]

You are given the job of making a model to predict the number of articles produced by a PhD student in their initial two years of PhD. You are given a dataset which contains four independent variables (Female, MentorArticles, Married, Children) and one dependent variable (Articles). These variables are defined as follows.

Articles: Number of articles published during the first two years of Ph.D.
Female: 1 if female scientist; 0 if male scientist.
MentorArticles: Number of articles published by the scientist mentor during the last 3 years.
Married: 1 if married; 0 otherwise.
Children: Number of children 5 or younger.

Let us assume this dataset contains 10 entries as shown in the table below.

Female	MentorArticles	Married	Children	Articles
1	32	1	1	2
0	20	0	0	1
1	17	1	0	0
1	21	0	0	0
0	49	1	2	0
0	10	1	3	0
1	20	0	0	1
1	35	1	1	0
0	37	0	0	3
1	13	1	2	0

Based on this data (dependent & independent variables), What kind of Generalized Linear Model (GLM) are you going to use for this situation and why?

[1 mark] Zero Inflated Poisson (ZIP) model should be used in predicting count of Articles for a PhD student's initial two years.

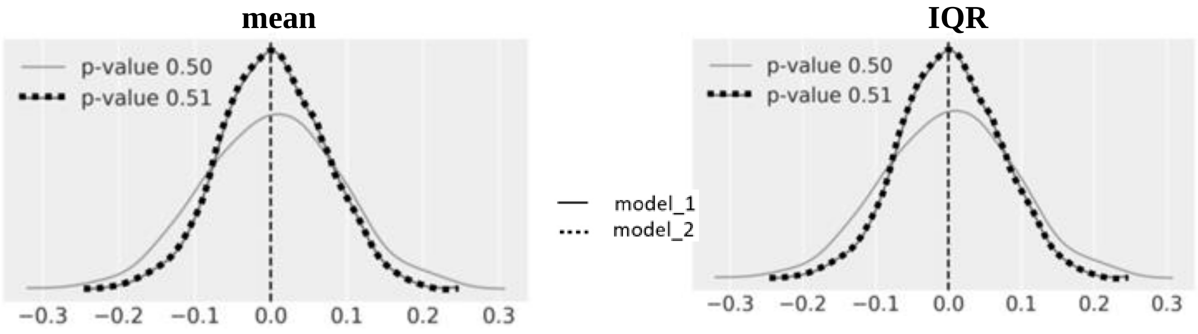
[2 marks] Reason: Given the nature of the dependent variable (Article Count) which is a count variable and there is potential excess zeros in the dependent variable. Poisson alone will not be suitable because of zero inflation. So ZIP should be used.

Question 5 [9+9+3= 21 marks]

Let's suppose we have created three models on a dataset, and we are interested in model selection from the available models that we have, in context of model comparison, we are interested to compare the models via following.

a. Comparison via PPC
b. Comparison via Information Criteria
c. Comparison via Bayes Factors

a) Comparison via PPC: Consider the following plot, next answer the questions


mean
IQR

i. [2+1 marks] Can we say the model_1 is in presence of a biased posterior predictive distribution? Your answer should be in **yes** or **no** also Justify in one line.

[1 mark] No,
[2 marks] Justification: According to the book, 'we should expect a p-value around 0.5; otherwise, we are in the presence of a biased posterior predictive distribution.' So model_1's p-value (0.50) indicates an unbiased posterior.

ii. [2+1 marks] Can we say the model_2 is in presence of a biased posterior predictive distribution? Your answer should be in **yes** or **no** also Justify in one line.
[1 mark] No,

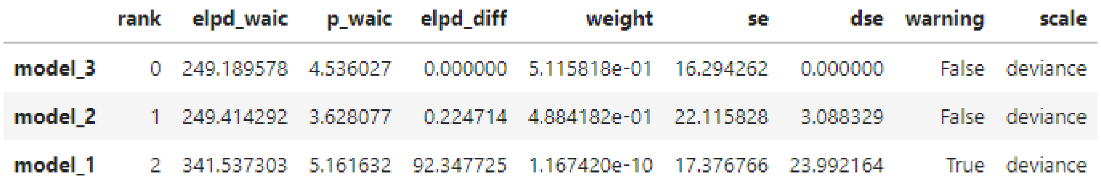
[2 marks] **Justification**: According to the book, 'we should expect a p-value around 0.5; otherwise, we are in the presence of a biased posterior predictive distribution.' So model_2's p-value (0.51) also indicates an unbiased posterior.

iii. [2+1 marks] Can we say both models are in the presence of a biased posterior predictive distribution? Your answer should be in **yes** or **no** also Justify in one line.
[1 mark] No

[2 marks] **Justification**: According to the book, Model_1's p-value (0.50) and Model_2's p-value (0.51) also indicate an unbiased posterior, as both are close to 0.5.

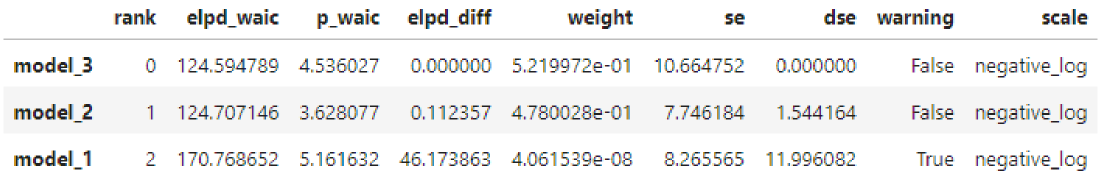
b) Comparison via Information Criteria: Consider the following three tables with different scales i.e. **log**, **deviance** and **negative log** next answer the questions.

Table 1



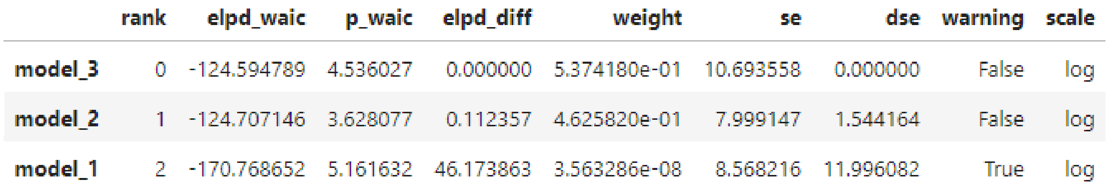
	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
model_3	0	249.189578	4.536027	0.000000	5.115818e-01	16.294262	0.000000	False	deviance
model_2	1	249.414292	3.628077	0.224714	4.884182e-01	22.115828	3.088329	False	deviance
model_1	2	341.537303	5.161632	92.347725	1.167420e-10	17.376766	23.992164	True	deviance

Table 2



	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
model_3	0	124.594789	4.536027	0.000000	5.219972e-01	10.664752	0.000000	False	negative_log
model_2	1	124.707146	3.628077	0.112357	4.780028e-01	7.746184	1.544164	False	negative_log
model_1	2	170.768652	5.161632	46.173863	4.061539e-08	8.265565	11.996082	True	negative_log

Table 3



	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
model_3	0	-124.594789	4.536027	0.000000	5.374180e-01	10.693558	0.000000	False	log
model_2	1	-124.707146	3.628077	0.112357	4.625820e-01	7.999147	1.544164	False	log
model_1	2	-170.768652	5.161632	46.173863	3.563286e-08	8.568216	11.996082	True	log

Comparison via LOO or WAIC

i. [3 marks] From the above tables select one model out of three which is better from the **predictive accuracy point** of view. Clearly mention which metric did you used for your comparison.

[1 mark] In all of the three tables above, model_3 is better than the other two from predictive accuracy point of view.
[2 marks] From the above three table **elpd_waic** used as metric of comparison.

ii. [3 marks] Review the three tables above and pick the model that is the **simplest** among model_1, model_2, and model_3. Clearly mention which metric did you used for your comparison.

[1 mark] In all of the three tables above, model_2 is simplest than the other two.
[2 marks] From the above three tables, **p_waic** used as metric of comparison.

iii. [3 marks] Now you need to select one model that is good from the above models in three tables W.r.t: **predictive accuracy** and **simplicity**. Choose one model out of three that has better predictive accuracy as well as simplicity. Clearly mention which metric did you used for your comparison.

[1 mark] model_2 is better than other two models w.r.t predictive accuracy and simplicity.
[2 marks] Based on elpd_waic values which are almost same for model_2 and model_3 which implies for accuracy point of view, more or less same. But from the simplicity point of view, model_2 is simpler than model_3 (based on p_waic)

c) Comparison via Bayes Factors: Consider the statistical models_1 i.e., model_2, model_3. Next the Bayes Factor (ratio of two marginal likelihoods) below in table.

$$\mathrm{BF}_1 = \frac{p(y \vee mode_3)}{p(y \vee model_1)} \quad , \quad \mathrm{BF}_2 = \frac{p(y \vee mode_2)}{p(y \vee model_3)} \quad , \quad \mathrm{BF}_3 = \frac{p(y \vee mode_2)}{p(y \vee model_1)}$$

BF_1	BF_2	BF_3
60	2	48

i. [1 mark] Consider all the BF_1 & BF_3 values. Can we say model_3 & model_2 both are better than model_1?

[1 mark] Yes:
With BF_1= 60 > 1 we can say model_3 with very strong evidence favoring model_3 than model_1
With BF_3= 48 >1 we can say model_2 with very strong evidence favoring model_1

ii. [1 mark] Based on BF_1 value, compare which model is better, i.e. model _3 or mode_1?

[1 mark] With BF_1= 60 > 1 we can say model_3 is better, with very strong evidence favoring model_3 than model_1

iii. [1 mark] Based on BF_3 value, compare which model is better, i.e. model _2 or mode_1?

[1 mark] With BF_3= 48 > 1 we can say model_2 is better, with very strong evidence favoring model_2 than model_1

Extra Page