National University of Computer and Emerging Sciences

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

MT-2002: Statistical Modeling

Saturday, 27th May, 2023

Course Instructors

Mukhtar Ullah, Shahzad Saleem, Muhammad Almas Khan

Serial No:

Final Exam

Total Time: 3 Hours

Total Marks: 90

_____

Signature of Invigilator

Solution

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 question and part number etc.
4. After asked to commence the exam, please verify that you have **twelve (12)** 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	**20**	**20**	**20**	**10**	**20**	**90**

Question 1 [20 Marks]

The Iris dataset is often used in machine learning and statistical modeling experiments. It contains measurements of various features of three different species of Iris flowers: Setosa, Versicolor, and Virginica. This dataset consists of 150 samples. The following table illustrates the structure of the iris dataset.

Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
.	.	.	.	.
.	.	.	.	.
6.9	3.1	5.4	2.1	Virginica

Assume you are performing experiments with two statistical models given below i.e., **Model 1** and **Model 2**.

$\alpha \sim$ Normal(mu, sigma)
$\beta \sim$ Normal (mu, sigma)
$\epsilon \sim$ HalfNormal(sigma)
$\mu =$ Deterministic $(\alpha + \beta * x)$
y_pred $\sim$ Normal(mu=μ, sigma = ϵ)

Model 1

$\alpha \sim$ Normal (mu, sigma)
$\beta_1 \sim$ Normal (mu, sigma)
$\beta_2 \sim$ Normal (mu, sigma)
$\epsilon \sim$ HalfCauchy(beta)
$\mu =$ $\alpha + \beta 1 * x + \beta 2 * (x**2)$
y_pred $\sim$ Normal (mu=μ, sigma = ϵ)

Model 2

Answer the following questions.

a) Suppose you are predicting "*Petal Length*" of iris dataset using "*Petal Width*" as predictor variable. Both variables are continuous. Give the pros and cons of the above-mentioned two models, i.e., **Model 1** and **Model 2,** in the context of this prediction. **[7]**

Answer: Model 1: Linear Regression:

Pros:

linear regression can capture linear relationships easily so if above two variables have linear relationship, then this model is good.

It is computationally efficient.

It can provide a better fit to the data compared to polynomial regression when the underlying relationship is linear. Linear regression can have improved predictive power by capturing linear patterns.

Cons:

Not best if above two variables have nonlinear relationships and if the relationship between the predictor and response is highly non-linear, a linear regression model may underfit the data and provide poor predictions.

Model 2: Polynomial Regression:

Pros:

polynomial regression can capture nonlinear relationships easily so if above two variables have nonlinear relationship, then polynomial regression is good.

It can provide a better fit to the data compared to linear regression when the underlying relationship is non-linear. Polynomial regression can have improved predictive power by capturing non-linear patterns such as curve, bends and turning points.

Cons:
Not best if above two variables have nonlinear relationships and including high-degree polynomial terms can lead to overfitting, where the model fits the training data too closely and performs poorly on new data.

b) Suppose for predicting the "*Petal Width*" one uses "*Petal Length*", "*Sepal Length*", and "*Sepal Width*" as predictors, Hence, now you have three predictors and one predicting variable.

Petal Length (cm)	Sepal Length (cm)	Sepal Width (cm)	Petal Width (cm)

Predictors *Predicting variable*

Suggest a statistical model by applying some modifications to **Model 1** or **Model 2** above. Write your modified model using statistical language. **[7]**

ANSWER: The multiple linear regression model is best choice for the above prediction problem. We can modify the model 1 from (a) which is simple linear regression in order to achieve the above prediction goal by including all other variables. we need to change the μ by including other variables and β coefficients.

α ~ Normal (mu, sigma)
β_1 ~ Normal (mu, sigma)
β_2 ~ Normal (mu, sigma)
β_3 ~ Normal (mu, sigma)
ϵ ~ HalfNormal(sigma)
μ = Deterministic ($\alpha + \beta_1$ * Petal_Length + β2 * Sepal_Length + β3 * Sepal_Width)

Petal Width_pred ~ Normal (mu=μ, sigma = ϵ)

c) Consider the scatter plot as shown in Figure 1 below between "*Petal Width*", and "*Petal Length*". Suppose you changed your model from **Part a above**, that resulted in a new scatter plot between "*Petal Width*", and "*Petal Length*" as shown in Figure 2. Suggest a new statistical model that caters for this change. **[6]**

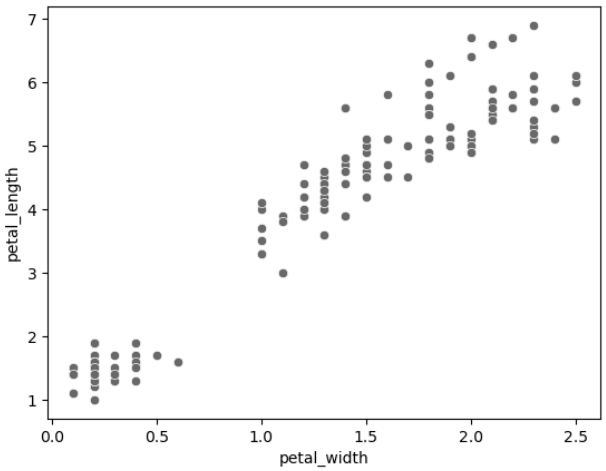


Figure 1

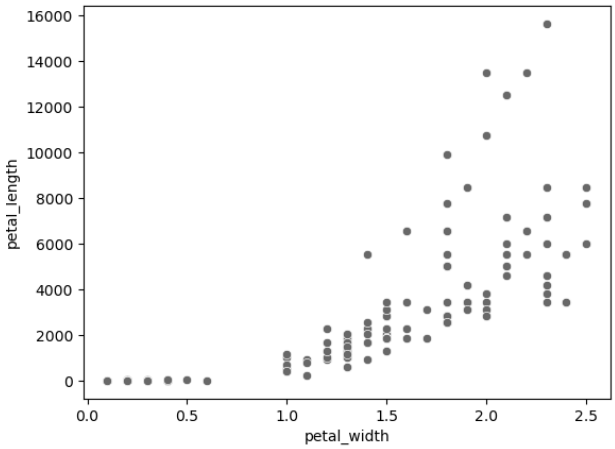


Figure 2

ANSWER: After the mathematical tricks the relationship between the data i.e two variables are now nonlinear so changing the models 1 in (a) will look like the following.

α	~ Normal (mu, sigma)
β_1	~ Normal (mu, sigma)
β_2	~ Normal (mu, sigma)
ϵ	~ HalfCauchy(beta)
μ	= Deterministic (α + β1 * *Petal Width* + β2 * (*Petal Width* **2)
Petal Length _pred	~ Normal (mu=μ, sigma = ϵ)

Question 2 [20 Marks]

Considering the iris dataset once again, answer the following questions.

a) Suppose we select "*Petal Width*" as predictor to predict the flower species. Considering the Model 1 and Model 2 given in Question 1, which model should be generalized to accomplish this task. Also write this new version of statistical model in statistical language.
[7]

ANSWER: Now the predicting variable here is discrete so few things will need to be change in order to achieve the above goal of flower classification.

1) Likelihood
2) Including sigmoid function

The above model i.e., Linear regression in (a) can be written as.

α ~ Normal (mu, sigma)
β ~ Normal (mu, sigma)
ϵ ~ HalfNormal(sigma)
μ = Deterministic ($\alpha + \beta$ * *Petal Width*)
θ = Deterministic (Sigmoid(μ))
Species_pred ~ Bernoulli (θ, observed data = *species*)

b) Now suppose one predictor is not sufficient for predicting the two species of flower. Thus, we pick all predictors, i.e., "*Petal Width*", "*Petal Length*", "*Sepal Length*", and "*Sepal Width*", to predict the flower species. Write the related statistical model in statistical language. **[7]**

ANSWER: As mentioned above, we will modify the model in (a) in such a way to include all predictor variables, for including all variable we need to change the μ and boundary decision. Consider the following modified model.

α ~ Normal (mu, sigma)
β_1 ~ Normal (mu, sigma)
β_2 ~ Normal (mu, sigma)
β_3 ~ Normal (mu, sigma)
β_4 ~ Normal (mu, sigma)
ϵ ~ HalfNormal(sigma)
μ = Deterministic ($\alpha + \beta_1$ * Petal_Length + β2 * Sepal_Length + β3 * Sepal_Width + β_4 * Petal_Length)
θ = Deterministic (Sigmoid(μ))
Species _pred ~ Bernoulli (θ, observed data = Species)

c) Suppose you want to predict more than two species of the flower using all predictor/independent variables, suggest a statistical modeling approach and write the model definition in statistical language. **[6]**

To predict more than two types of flowers we need to change the following things.
1) we need to include Softmax function
2) we need to change the likelihood.
The following is the modified version of statistical model that will handle the above problem.

α ~ Normal (mu, sigma)
β_1 ~ Normal (mu, sigma)
β_2 ~ Normal (mu, sigma)
β_3 ~ Normal (mu, sigma)
β_4 ~ Normal (mu, sigma)
ϵ ~ HalfNormal(sigma)
μ = Deterministic ($\alpha + \beta_1$ * Petal_Length + β2 * Sepal_Length + β3 * Sepal_Width + β_4 * Petal_Length)
θ = Deterministic (Softmax(μ))
Species _pred ~ Categorical (θ, observed data = Species)

Question 3 [20 Marks]

a) Suppose a team of researchers is studying the population dynamics of iris flowers in a specific region. They collect data on the species of iris flowers found in a designated area during weekly surveys. In addition to the standard weekly average measurements of "*Petal Width*", "*Petal Length*", "*Sepal Length*", and "*Sepal Width*", they also record the "*Count*" of iris flowers observed during each survey. The

Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Species	Count
5.1	3.5	1.4	0.2	Setosa	10
4.9	3.0	1.4	0.2	Setosa	11
.	.	.	.	.	
.	.	.	.	.	
6.9	3.1	5.4	2.1	Virginica	20

modified dataset would look like the following.

Suppose you perform an experiment with this new version of iris dataset, in which you'll use all four predictors, i.e., "*Petal Width*", "*Petal Length*", "*Sepal Length*", "*Sepal Width*", to predict the "*Count*" variable. Suggest a regression model in this case and write it in statistical language. **[7]**

Based on the above prediction task as stated, the recommended model is Poisson regression model. Since count variable contains only the count of flowers very less number of zero are there in count variable.

α ~ Normal (mu, sigma)
β_1 ~ Normal (mu, sigma)
β_2 ~ Normal (mu, sigma)
β_3 ~ Normal (mu, sigma)
β_4 ~ Normal (mu, sigma)
ϵ ~ HalfNormal (sigma)
μ = Deterministic ($\alpha + \beta_1$ * Petal_Length + β2 * Sepal_Length + β3 * Sepal_Width + β_4 * Petal_Length)
θ = Deterministic (exp $^{\mu}$)
Species _pred ~ Poisson (θ, observed data = Species)

b) Consider logistic regression (LR) and linear discriminant analysis (LDA) models. Suppose the accompanied dataset has three categories and four predictors. In the light of this information, what type of values will be witnessed in the posterior predictive of both the LR and LDA. **[4]**

When comparing the posterior predictives, the logistic model is binary, estimating either 0 or 1, while the LDA model has real numbers. The logistic regression makes predictions as to which class (category) a particular predictor belongs to, whereas the LDA model makes predictions about the predictors directly.

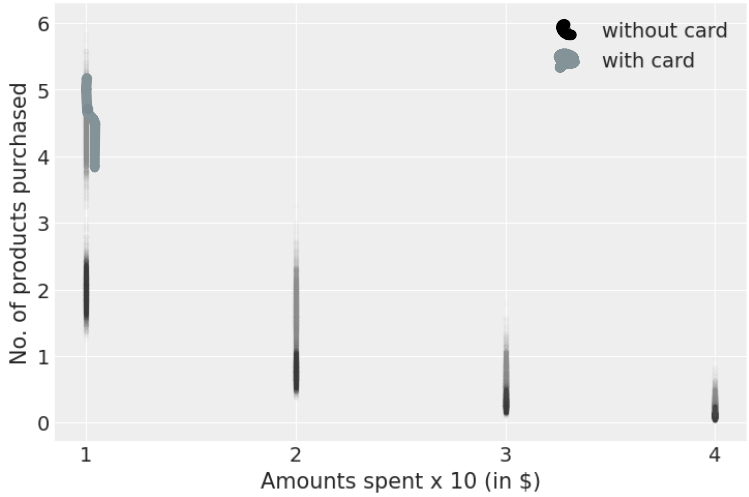
c) Consider a dataset of 250 customers purchasing from a website. Here are some parts of the data per customer:

- Number of products one purchases (*count*)
- Amounts spend (in $) by each customer (*amounts*)
- Whether a customer holds a credit/debit card or not (*card*)

Using this data, write a ZIP-regression model using statistical language that predicts the number of products purchased as a function of the amounts spent and card variables.
Hints: The inverse link function to be used in this case is $\theta = e^{(\alpha + X\beta)}$. Also remember that ZIP-regression model always features a prior Ψ which is usually declared as a beta(1, 1) distribution. **[6]**

ϕ ~ Beta(α, β)
α ~ Normal(μ, σ)
β ~ Normal(μ, σ, shape=2)
θ = exp(α + β[0] * data['amounts'] + β[1] * data['card'])
y ~ ZIP(ψ, θ, observed= data['count'])

d) Consider the following plot related to the dataset in **Part c above**. What can be inferred from this plot regarding model outcomes? **[3]**



From above *Figure*, for each additional $10 spent, the expected no. of products purchased decreases. Also, people who have a credit/debit card. purchased more products.

Question 4 [10 Marks]

The following table shows the results of comparison of three models with ArviZ.

	waic	waic_se	p_waic	loo	loo_se	p_loo	scale
Model_1	9.13	5.35	2.67	8.92	5.24	2.56	deviance
Model_2	28.77	4.62	2.46	28.72	4.53	2.45	deviance
Model_3	8.77	8.62	2.96	8.72	6.53	2.86	deviance

Now answer the following questions:

a) Which model is the simplest of all? Justify your answer by referring to the appropriate columns. **[4]**

Model_2 is the simplest of all based on the smallest values in columns p_waic

b) Which model is better of all from predictive accuracy point of view? **[3]**
Justify your answer by referring to the appropriate columns.

Model_3 is better because of smallest values in columns waic and loo representing WAIC and LOO, respectively.

c) Which model is the best of all, from predictive accuracy & complexity point of view? **[3]**
Justify your answer by referring to
the appropriate column.

Since WAIC and LOO both incorporate generalization performance i.e predictive accurcuracy, the answer is the same as in part b.

Model_3 is the best of all based on the smallest values in columns waic and loo representing WAIC and LOO, respectively, p_waic comparativly less high in comparison to other models.

Question 5 [20 Marks]

For each of the following, tick (✔) the correct choices and cross (✘) the incorrect ones..

1) When computing Bayes factors,

- ☒ more samples should be drawn from the better model.
- ☑ both models should be visited equally for sampling.
- ☑ unrestricted parameter drifting can be solved by using pseudo priors.
- ☑ marginal likelihood can be estimated from sequential Monte Carlo.

2) Information Criteria are preferred over Bayes factor because

- ☒ Bayes factors are not sensitive to priors.
- ☒ Information Criteria are focused on which model is better.
- ☒ Bayes factors are focused on which model will give the better predictions.
- ☒ Priors are directly used in computation of information criteria.

3) Regularization of priors

- ☑ often takes the form of penalizing larger values for the parameters in a model.
- ☒ introduces a bias to increase generalization error.
- ☒ takes the form of ridge regression if variable selection is desired.
- ☑ takes the form of Lasso regression if a sparse model is desired.

4) The widely applicable information criterion (WAIC)

- ☒ is a combination of log point-wise predictive density (lppd) and deviance.
- ☑ employs variance of the log-likelihood over posterior to correct the overestimation of lppd.
- ☑ penalizes the spread of the posterior.
- ☑ penalizes the number of effective parameters.

5) Entropy

- ☒ of a Gaussian prior is larger than that of a flat prior
- ☒ is computed from logarithms of possible values of a random variable
- ☑ can be used to justify a regularizing prior
- ☑ can be used to justify a weakly informative prior