

United Arab Emirates University
STAT 380
Midterm Exam

Name:

ID:

- There are a total of 110 points in this Question Paper. Answer as much as you can. If your acquired score is greater than equal to 100 it will be counted as 100%.

- There are three parts in this Exam. Part-I involves TRUE/FALSE or multiple choice answer type questions, Part-II contains a few short answer type questions, whereas Part-III consists of one descriptive answer type question.

- The Exam is scheduled for 75 minutes

For instructor's use only

Problem Number	Obtained Score	Total Score
Problem 1		45
Problem 2		25
Problem 3		20
Problem 4		20
TOTAL		110
TOTAL(out of 100)		100

Part-I

Pick the correct answer option for the questions in this part of the exam.

Consider a dataset with 3 covariates (explanatory variables) to model a continuous response variable. 2 of the 3 covariates are numerical in nature while one of the covariates, 'Highest Education level' is a categorical variable that can take either of the following THREE values:

{ High-School, Bachelor's degree, Master's Degree or higher }

1. (a)

Note that, it is customary to introduce appropriate (multiple) 'dummy'-variables as co-variates to incorporate categorical covariates. How many total regression coefficients, including the intercept, are there in the constructed model ?

Score:
Total Score: 5

Ans:

One of the primary assumptions of the standard linear regression is the Normality of the corresponding errors. To check the Normality of the Model errors, a practitioner conducted the standard Shapiro-Wilks's test for testing the Normality of the corresponding residuals and obtained a p-value of 0.0034. Identify whether the following statement is TRUE or FALSE.

(b)

Statement: Based on the results from the Shapiro-Wilks's test on the model residuals, we have strong statistical evidence that the corresponding model errors are not Normally distributed.

Score:
Total Score: 5

Ans:

In the case, when performing the Standard Linear Regression on a dataset to model a continuous response, a statistical analyst wants to verify whether the assumption of the independence of the model errors is satisfied or not. Which one from the following statistical hypothesis testing procedure is the most appropriate for the purpose?

(c)

Score:
Total Score: 5

Ans:

If we consider a 3-degree regression splines with 5 knot points, then how many regression coefficient parameters are there in the model? (i.e. , what is the dimension of the model?)

(d)

Score:
Total Score: 5

Ans:

(e)

Let $\pi \in (0,1)$ be such that $\text{Logit}(\pi) = -1.1$. Choose the most appropriate value of $\text{Odds}(\pi)$ from the list below.

Score:
Total Score: 5

Ans: 0.333 3.004 0.249 0.082

(f)

Using a statistical machine-learning procedure, it is estimated that the probability of at least 10 years of "survival" of a randomly chosen patient who went through a specific medical procedure is 0.75. Based on the information, calculate the odds for at least 10 years of "survival" of the patient ?

Score:
Total Score: 5

Ans: 0.25 3.0 4.0 1.098

(g)

What is the value of the following function $\text{Logit}(0.75)$?

Score:
Total Score: 5

Ans: 0.679 1.099 2.117 0.472

(h)

Consider analyzing a dataset that has a binary categorical response while all the covariates are continuous (numerical) in nature. We know both the models; logistic regression and the Quadratic Discriminant Analysis (QDA) can be applied to the corresponding classification problem. To compare their performance, the corresponding ROC curves are constructed. The resulting Area Under the ROC Curve (AUC) is calculated based on a Testing set. **The AUC for the logistic regression is obtained to be 0.89 while the AUC for the QDA appears to be 0.95.** Identify whether the following statement is TRUE or FALSE.

Statement: Based on the AUC criterion, the performance of Logistic Regression is better than that of the QDA for this data set.

Score:
Total Score: 5

Ans: TRUE FALSE

(i)

If we consider a Dataset that has 130 observations and 210 covariates (independent variables) to model a continuous response variable. Identify whether the following statement is True or False.

Statement: It is **Not Possible** to fit a Standard Linear Regression as the dataset can be considered to be high-dimensional data.

Score:
Total Score: 5

Ans: TRUE FALSE

Part-II

Answer the following short type questions. Show your steps to get full credit.

2.

A newly developed spam-filtering algorithm is implemented in all the email user accounts of a corporate office. Based on a **total of 1354 external emails** received in the first few days, the company summarized the following data to evaluate its performance.

Among the 1354 emails that are considered it appears that in actuality (TRUTH), there is a total of 271 spam emails while the rest of the external emails are not spam. The Spam-filtering algorithm predicts and labels an external email to be either a Spam ('Positive' for spam) email or a 'Not-Spam' ('Negative' for Spam'). However, the Algorithm is not always accurate.

Out of the **271 spam emails the algorithm can correctly detect only 233**. On the other hand, it correctly identifies a total of **1071 out of 1083 non-Spam emails**. Based on the provided information, answer the following questions.

Construct a classification table for assessing the performance of the 'spam-filtering algorithm'.

(a)

Score: _____
Total Score: 10

Calculate the 'Sensitivity' of the spam-filtering algorithm in detecting a spam email.

(b)

Score: _____
Total Score: 5

Evaluate the 'Specificity' of the spam-filtering algorithm in identifying a not-spam email.

(c)

Score: _____
Total Score: 5

What is the value of the corresponding Yuden-Index?

(d)

Score: _____
Total Score: 5

This problem pertains to the dataset on the O-Ring failure of the Space Shuttles. It was known that there is an association between the O-Ring seal failure and the low atmosphere temperature during the corresponding shuttle launch. The variable "oringFail" in the data set indicates whether the shuttle experienced an O-ring failure during its launch. The "temperature" column lists the outside temperature at the time of the shuttle launch. A logistic regression model is considered with the following specification:

3.

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\text{Logit}(\pi_i) := \beta_0 + \beta_1 \times \text{'temperature'}.$$

Here the response variable $Y_i = 1$ if there is a 'O-Ring' failure corresponding to the i^{th} data point. Answer the parts of this questions based on the following output from the R Statistical Software that is provided below:

Call:

```
glm(formula = oringFail ~ temperature, family = "binomial", data = oring12)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2034	-0.7444	-0.4970	0.3563	2.0059

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	10.18873	5.17679	1.968	0.0491 *
temperature	-0.16076	0.07457	-2.156	0.0311 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34.795 on 29 degrees of freedom

Residual deviance: 28.688 on 28 degrees of freedom AIC: 32.688

(a)

Interpret the estimated value of regression coefficient corresponding to the variable 'temperature' in the context of the specific problem.

Score: _____
Total Score: 10

(b)

Based on fitted model, derive the predicted probability of O-ring failure if the corresponding 'temperature' is 40 degrees Fahrenheit?

Score: _____
Total Score: 10

Part-III

Answer the following descriptive type questions. Show your steps to get full credit.

4. Let us consider a Dataset that has a continuous response Y , and numerical continuous covariates $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$. The observed data is provided as $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, where n is the number of observations. Assume that $n \geq 100$.

- (a) Write down the objective function of a Ridge Regression model where denote the corresponding model selection parameter (tuning parameter) to be λ .

Score: _____
Total Score: 5

- (b) For a given value of $\lambda > 0$, what is the formula for $\hat{\beta}_{\sim \text{Ridge}}$, the corresponding estimated regression coefficients for the regression parameter β ?

Score: _____
Total Score: 5

(c)

Write down the details of the Cross-Validation procedure to select the optimal value for the model selection parameter $\lambda > 0$.

Score: _____
Total Score: 10