Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN STATISTICAL SCIENCE

END OF SEMESTER EXAMINATION

STA 8303: STATISTICAL DATA MINING

Date: August 23, 2023 TIME: 3 Hours

Instructions

1. This examination consists of **FIVE** questions.
2. Answer question **ONE** and any other **TWO** questions

Question 1 (20 Marks)

a) The ordinary least squares (OLS) estimates are obtained by minimizing the residual sum of squared for a model of the form

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a $(n \times 1)$ vector of response value, $\mathbf{X}$ is a $(n \times p)$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ error vector.
A basic requirement to perform ordinary least squares regression (OLS) is that the inverse of the matrix $\mathbf{X}'\mathbf{X}$ exists. However, in certain situations $(\mathbf{X}'\mathbf{X})^{-1}$ may not be calculable.
Suggest at least two instances where this may be the case.

(**5** Marks)

b) The conditional number (CN) and the variance inflation factor (VIF) are metrices used to diagnose one of the problems in part (a) above. Explain what problem this is and how VIF and CN can be used to detect and diagnose this problem.

(**6** Marks)

c) Ridge regression, Lasso regression and elastic net regression are 3 approaches used in combating *Multicollinearity* in data. Distinguish between them, explaining advantages of each technique. In each case, provide an expression for the cost function that forms the basis for parameter estimation

(**9** Marks)

Question 2 (20 Marks)

a) In statistical learning, distinguish between supervised and unsupervised learning. Give appropriate examples of methods that fall into each of these categories.

(**6** Marks)

b) Explain how the each of the following resampling techniques is implemented in predictive modeling:
   i) Validation set approach.
   ii) Leave-One-Out cross-validation.
   iii) Bootstrapping.

(**8** Marks)

c) Given a random sample of size n, $X_1, X_2, \ldots, X_n$ from a population with probability density function $f(x, \theta)$. Explain how you would determine a 95% bootstrap confidence interval for the parameter θ.

(**6** Marks)

Question 3 (20 Marks)

a) **Prediction accuracy** and **Interpretability are two reasons why** data analysts are often not satisfied by the ordinary least squares estimates.
   i) Explain what these two terms mean.
   ii) Explain how these 2 issues can be resolved and suggest any standard techniques that can be used to improve on the limitations posed by these two problems.
   iii) Of the 2 approaches proposed in part (iii), mention any drawbacks you are aware of.

(**7** Marks)

b) Ridge regression offer a solution to the problem of prediction accuracy and interpretability associated with ordinary least squares regression. Supposed that we have data $(\mathbf{x}_i, y_i), \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and y_i are the regressors and response for the i −th observation. We also assume that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is the relationship relating **y**, the vector of outcome variables, and **X**, the design matrix of the predictors, and where $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are the vector of parameters and residuals, respectively.
   i) Given a penalty term λ, give an expression, in matrix form, for the cost function used to determine $\widehat{\boldsymbol{\beta}}_{Ridge}$, the ridge regression estimators of $\boldsymbol{\beta}$.
   ii) Minimize the cost function given in part(b) and show that $\widehat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$
   iii) Determine the expected and variance of this estimator. Comment on the Bias and efficiency of this estimator in comparison with the OLS estimator.

(**13** Marks)

Question 4 (20 Marks)

a) Logistic regression, Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are three classification techniques that are widely used by predictive modelers.

Explain the main similarities and differences that exist between LDA and QDA. Provide a mathematical description of each approach.

(**7** Marks)

b) Consider a data set of 144 observations of household cats. The data contains the cats' gender, body weight and height. The researcher would like to model and accurately predict the gender of a cat based on previously observed values.
To verify and test our model's performance, they split the data into training (60%) and test sets (40%). Two models were entertained:

- Model 1: A logistic regression model with body weight as predictors
- Model 2: A logistic regression model with body weight and height as predictors

The results of these two models are presented in Table 1 and Table 2. The confusion matrices for these two models are also presented in

Table 1 The results of fitting a logistic regression model with body weight as predictor to the training data (Model 1)

```
Call:
glm(formula = Sex.f ~ Bwt, family = binomial, data = training)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.7939     1.8571  -3.658 0.000254 ***
Bwt           2.8989     0.7346   3.946 7.94e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 111.559  on 87  degrees of freedom
Residual deviance:  89.159  on 86  degrees of freedom
AIC: 93.159
```

Table 2 The results of fitting a logistic regression model with body weight and height as predictors to the training data (Model 2)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8330     1.8334  -3.727 0.000194 ***
Bwt           3.5369     1.1111   3.183 0.001457 **
Hwt          -0.1602     0.2021  -0.792 0.428095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 111.559  on 87  degrees of freedom
Residual deviance:  88.527  on 85  degrees of freedom
AIC: 94.527
```

Table 3 Confusion matrix for Model 1 and 2

Actual status	Predicted status: Female	Predicted status: Male
Female	12	10
Male	13	22

Actual status	Predicted status: Female	Predicted status: Male
Female	9	15
Male	13	20

i) From the confusion matrices above, compare the 3 models. Compare your results based on model accuracy.

[**5** Marks]

ii) For the best fitting model, compute the following measures: sensitivity, specificity and the false positive rate.

[**8** Marks]