Strathmore
UNIVERSITY

INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN STATISTICAL SCIENCES
END OF SEMESTER EXAMINATION
STA 8203: PREDICTIVE MODELING AND STATISTICAL LEARNING

DATE: 22nd April 2024 Time: 3 Hours

Instructions

1. This examination consists of **FIVE** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1 (20 Marks)

a) Explain what EDA is.
   i) Distinguish between EDA, classical and Bayesian analysis
   ii) Enumerate the EDA assumptions

(**6** Marks)

b) The Box-cox transform can be used to remove skewness in data. Describe this approach and also explain how maximum likelihood estimation can be used to estimate the transformation parameter λ.

(**6** Marks)

a) Suppose that the probability density function $f(y,\theta)$ of a random variable Y belongs to the exponential dispersion family. Thus $f(y;\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right]$, where $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions, and the range of Y does not depend on θ or ϕ. We also assume that the distribution is parameterized in terms of the mean of Y, μ so that $\theta \equiv g(\mu)$ for some function g, then $g(\mu)$ is the canonical link. Show that:
   i) $E(Y) = b'(\theta)$
   ii) $Var(Y) = a(\phi)b''(\theta)$

(**8** Marks)

Question 2 (20 Marks)

a) Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim MVN(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.
   i) Explain what the hat-matrix is.
   ii) Explain how standardized residuals are used in regression diagnostics and use a mathematical approach to show that

$$Z = \frac{e_i}{s.e.(e_i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \sim N(0,1)$$

   iii) Explain how outliers, high-leverage values, and influential observations on the basis of the hat-matrix.

(**12** Marks)

b) The Dixon and the generalized (extreme Studentized deviate) ESD (Rosner) tests are approaches used in exploratory data analysis. Distinguish between them and explain in (mathematical) detail how each approach works.

(**6** Marks)

Question 3 (20 Marks)

a) Given a random sample $Y_1, \ldots, Y_n$ of size n from

$$f(y;\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right],$$

where $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions, and the range of Y does not depend on θ or ϕ.
Let $\boldsymbol{X} = (X_1, \ldots, X_k)'$ be a $(k \times 1)$ vector of covariates with the systematic component as

$$\eta = g(\mu_i) = \beta_1 X_{1i} + \cdots + \beta_j X_{ji} \ldots + \beta_k X_{ki}$$

Show that the estimating equation can be expressed as:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \times \frac{x_{ji}}{g'(\mu_i)} = 0,$$

where ℓ is the log-likelihood function.

[**12** Marks]

b) Consider a data set of 1,000 men aged 15-49 years screened for tuberculosis (TB). The data contains the patient's gender, BMI and HIV status. The researcher would like to model and accurately predict the TB status (positive or negative) of patients 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 **HIV status as predictors**
- Model 2: A logistic regression model with **gender, BMI and HIV status as predictors**

The confusion matrices for these two models are also presented in

Table 1 Confusion matrix for Model 1 and 2

	Predicted status	
Actual status	Positive	Negative
Positive	650	80
Negative	70	200

	Predicted status	
Actual status	Positive	Negative
Positive	500	100
Negative	100	300

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

[**4** Marks]

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

[**4** Marks]

Question 4 (20 Marks**)**

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

(**5** Marks)

b) The variance-bias trade-off is an important consideration in predictive modeling. Explain why and derive an expression for the mean-square error of vector of parameters

(**7** marks)

c) Concerning bias-variance decomposition.
   i) Provide a sketch of typical (squared) bias, variance, training error, and test error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.
   ii) Explain why each of the four curves has the shape displayed in part (a).

(**8** marks)

Question 5

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)