



Strathmore
UNIVERSITY

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

DATE: **December 17, 2021**

Time: **2 Hours**

Instructions

1. This examination consists of **FOUR** 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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I})$.
- Explain what the hat-matrix is.
 - Explain how standardized residuals are used in regression diagnostics and use a mathematical approach to show that

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

- 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.

[8 Marks]

Question 3 (20 Marks)

- a) Given a random sample $Y_1, \dots, Y_n$ 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 $\mathbf{X} = (X_1, \dots, X_k)'$ be a $(k \times 1)$ vector of covariates with the systematic component as

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

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 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 male 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 confusion matrices for these two models are also presented in

Table 1 Confusion matrix for Model 1 and 2

Actual status	Predicted status		Actual status	Predicted status	
	Female	Male		Female	Male
Female	12	10	Female	9	15
Male	13	22	Male	13	20

- From the confusion matrices above, compare the 3 models. Compare your results based on model accuracy.
- For the best fitting model, compute the following measures: sensitivity, specificity and the false positive rate.

[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 Validation set approach and Leave-One-Out cross-validation are resampling techniques implemented in predictive modeling. Describe each approach, enumerating its advantages and disadvantages.

[8 marks]

- c) 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]