INDIAN STATISTICAL INSTITUTE
Semester Examination 2025-26

M.Stat. - 1st Year
Regression Techniques

27 November, 2025

Maximum Marks: 100 Time: 3 hours

[Note: Each student is allowed to bring **two a4 papers** with hand-written notes. Notations are as used in the class. Answer as much as you can. Best of Luck!]

1. Consider a simple linear regression model $y = \alpha + \beta x + \epsilon$ with n observations (x_i, y_i) for $i = 1, \ldots, n$, and standard assumptions on the random error ϵ. Define b_{ij} to be the slope of the straight line joining the points (x_i, y_i) and (x_j, y_j) for $i, j = 1, \ldots, n$. Prove that the least-squares (LS) estimate of β can be expressed as a convex combination of $\{b_{ij} : i, j = 1, \ldots, n\}$. Interpret this result.

[6+2]=8

2. Consider a multiple linear regression model with all standard assumptions where the true mean function is given by

$$E(Y|X_1 = x_1, X_2 = x_2) = 3 + 4x_1 + 2x_2.$$

Derive the condition under which $E(Y|X_1 = x_1)$ is linear but has a negative coefficient for x_1. Discuss the implications of this result in the context of model building.

[5+2]=7

3. Consider a regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2[1 + (\gamma x_i)^2]) \quad \text{independently for} \quad i = 1, \ldots, n.$$

(a) Derive a set of consistent estimators for the parameters β_0, β_1, σ^2 and γ^2.

(b) Using part (a), or otherwise, develop a test for homogeneity of the error variances in the above-mentioned regression model.

[7+5]=12

4. Define the externally studentized residuals in the context of a multiple linear regression model, and use it to develop a statistical significance testing procedure to check if a particular observation is an outlier or not.

[1+4]=5

5. Consider fitting the following polynomial regression model

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2,$$

based on n observations (y_i, x_{1i}, x_{2i}) for $i = 1, \ldots, n$. Assume standard assumptions on the additive random errors in this regression model.

(a) Estimate the optimal (X_1, X_2) combination that maximizes the fitted response, clearly specifying the required condition (on the regression coefficients) that ensures the existence of such a unique maximizer.

(b) Find (asymptotic) standard errors of the estimates obtained in Part (a).

[(3+1)+(3+3)]=10

6. Suppose that observations $y_1, \ldots, y_n$ are independent with $y_i \sim$ Poisson(μ_i) for all $i = 1, \ldots, n$, where

$$\mu_i = \begin{cases} \alpha & i = 1, \ldots, m, \\ \alpha + \beta & i = m+1, \ldots, n. \end{cases}$$

(a) Derive the MLE for α and β in closed forms.

(b) Derive the deviance for this model and use it to develop a deviance based test for $H_0 : \beta = 0$.

[4+(4+4)]=12

7. Consider a multiple linear regression model $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$, where ϵ_is are independent and identically distributed with mean zero and variance σ^2 for $i = 1, \ldots, n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. Suppose we transform the responses (y_i) using the Box-Cox method with parameter λ.

(a) Derive the joint maximum likelihood estimator of $\boldsymbol{\beta}$, σ^2 and λ.

(b) Using part (a), or otherwise, develop a test for the linearity assumption in the original regression model of Y on $\boldsymbol{X}$ against the Box-Cox type alternatives.
[You need to write down both test statistic and critical value explicitly to get full marks]

[7+5]=12

8. Suppose that, given n IID observations (y_i, x_{1i}, x_{2i}) , $i = 1, \ldots, n$, on a random vector (Y, X_1, X_2), we wish to fit a non-parametric regression model given by

$$y_i = f(x_{1i}, x_{2i}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where f is unspecified and ϵ_i s are IID random noises with mean zero and variance σ^2.

(a) Write down f in terms of 2-dimensional (product) B-spline basis.

(b) Explain how one can estimate the underlying parameters (including σ^2) based on the observed data.

[4+6]=10

9. Suppose that, given the value of predictor $\boldsymbol{X} = \boldsymbol{x}$, the response Y is exponentially distributed with mean $\boldsymbol{x}^T\boldsymbol{\beta}$. Prove that, given n observations $(y_i, \boldsymbol{x}_i)$ from this model, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the same as it generalized LS (GLS) estimator.

[5]

10. Prove or disprove the following statements, with justifications:

(a) There exists an $n \times p$ design matrix $\mathbb{X}$ whose second and third rows are the same, and the second diagonal element of its hat matrix is 0.75.

(b) Multicollinearity is harmless if the objective of the analysis is prediction only.

(c) The cross-validation error is a better estimate of the true error than the training error.

(d) The exclusion of an important variable from a multiple linear regression model may produce a significant Durbin-Watson d value.

(e) In a linear regression model of Y on (X_1, X_2, X_3) under usual (standard) assumptions, its coefficient of determination can be expressed as $R^2 = r_{13}^2 + r_{12.3}^2$.

[4+4+3+4+5]=20

11. (Buffer question) Consider a penalize regression estimator based on n observations from a usual multiple linear regression model with p-dimensional predictors, which is defined as:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \qquad q \in (0, 2].$$

(a) Derive a formula for the j-th component of $\widehat{\boldsymbol{\beta}}_\lambda$ when q is fixed (given) and the design matrix $\mathbb{X}$ satisfies $\mathbb{X}^T\mathbb{X} = \mathbb{I}_p$. Can you relate this estimator to the corresponding Ridge and LASSO estimators?

(b) Using part (a), or otherwise, develop an algorithm for solving this penalized regression problem for a given $q \geq 1$ when $\mathbb{X}^T\mathbb{X} \neq \mathbb{I}_p$.

(c) Prove that $\widehat{\boldsymbol{\beta}}_\lambda$ corresponds to the best subset regression estimator if we set $q = 0$.

[(6+1)+4+3]=14