Indian Statistical Institute
B S D S, Second Year, First Semester, 2025-26
Final Examination
Statistics III: Multivariate Data and Regression

08.12.25 **Maximum Score 100** **Duration: 180 minutes**
Total score 110

Name **Student ID**

---

1. Write your name and ID on each page.
2. Numbers in brackets denote total points allotted to each question.
3. You may use calculator.
4. Laptops and phones are not allowed.
5. You are allowed to bring one page (2 sided) or 2 pages (1 sided) of notes. No other material is allowed.
6. Show all your work.

Name **Student ID**

---

1. (4+2+6=12) Suppose you have fitted two regression lines $Y = -0.5 + 2.5X$ and $Y = 0.4 + 1.6X$ on 40 pairs of observations. One line is the least squares regression line of Y on X and the other is the regression line of X on Y.

   (a) Identify which is the regression line of Y on X.

   (b) For the value 20 of X, what is the estimate of the expected value of Y?

   (c) Find the sample means of X and Y, the sample correlation coeffcient between the two variables and the ratio of the standard deviations of X and Y.

Name **Student ID**

---

2. (3+3+3+3=12) In each of the following situations, set up the model (multiple regression, ANOVA, logistic regression, chi-square test etc) identifying all the variables in the model with the physical problem.

   (a) Amount of credit availed (money taken on loan) by a random sample of individuals from a village, as a function of their gender, education level and caste.

   (b) Is there a difference in the amount of credit availed for different castes?

   (c) The chance of a person paying back the loan in time as a function of their gender, education level and caste.

   (d) Is there a difference in the chance of paying back loan for different castes?

Name **Student ID**

---

3. (10+2=12) Suppose the random variable Y comes from a true model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where ϵ has mean zero and variance σ^2. Suppose we have independent observations $(y_i, x_{1i}, x_{2i}), i = 1 \cdots, n$ where x_{1i}, x_{2i} are fixed. We fit a linear model of y on x_1 (only) by least squares to obtain the estimators $\hat{\alpha}$ and $\hat{\beta}_1$.

   (a) Find the expected value of $\hat{\beta}_1$ under the true model and hence find the bias in estimation of β_1.

   (b) Under what condition on the points x_{1i}, x_{2i} is $\hat{\beta}_1$ unbiased?

Name **Student ID**

---

4. (4+4+4+4=16) Consider the following model:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where Y and ϵ are n dimensional vectors, X_1 and X_2 are $n \times p$ and $n \times q$ dimensional matrices of predictors, β_1 and β_2 are unknown regression coefficient vectors of dimensions p and q.

Let SSR_1 be the sum of squares residuals from the model and SSR_2 be the sum of squares residuals when $\beta_2 = 0$. Assume that ϵ has a multivariate normal distribution with mean zero and variance $\sigma^2 I$.

(a) Show that SSR_1 follows a χ^2 distribution. What are the degrees of freedom?

(b) Under the hypothesis $H_0 : \beta_2 = 0$ show that SSR_2 follows a χ^2 distribution. What are the degrees of freedom?

(c) Under H_0, show that SSR_2-SSR_1 follows a χ^2_q distribution and is independent of SSR_1.

(d) Form an F statistic to test H_0.

Name **Student ID**

---

5. (4+4+4+4=16) Consider a logistic regression model with a single predictor and no intercept.
   (a) What is the log likelihood of the slope parameter β.
   (b) Write down the score equation, that is, find the derivative of the lig likelihood and equate it to zero.
   (c) Find the iterative equation of the Newton-Raphson method.
   (d) Find the iterative equation of the Fisher scoring method.

Name **Student ID**

---

6. (4+8=12) The iris data consists of 4 characters (sepal length, sepal width, petal length, petal width) measured on 50 flowers from each of 3 species (setosa, versicolor, virginica). We run the following command in R.

```
summary(aov(formula = Sepal.Width ~ Species, data = iris))
```

   (a) Complete the table of output.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species		11.35			<2e-16
Residuals			0.115		

   (b) Carry out the ANOVA test using the above output stating the null and alternative hypotheses, assumptions and conclusions.

Name **Student ID**

---

7. (12) Explain what the following R code and output is doing. The data is on hair and eye color of 592 individuals. State the model, hypotheses, data, assumptions, test statistic, its distribution and conclusion.

```
> data
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
> chisq.test(data)

Pearson's Chi-squared test

data:  data
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

Name **Student ID**

---

8. (4+4=8) Suppose you have a sample of eighth graders from a school and you measure their heights on Jan 1st (x) and Dec 31st (y) of the same year.

   (a) Draw a possible scatterplot to show that the heights of males and females have increased separately, but if you do a regression of y on x, the slope is negative.

   (b) How do you handle this situation in a regression analysis, if your objective is to predict the Dec 31st height for a student on Jan 1st?

Name **Student ID**

---

9. (2+4+4=10) You roll a die 100 times and Y_i denotes the number of times i occues, $i = 1, \cdots, 6$.
   (a) What is the joint distribution of $(Y_1, \cdots, Y_6)$?
   (b) What is the joint distribution of (Y_1, Y_2, Y_3)?
   (c) If you know that 6 is observed 20 times, then what is the joint distribution of (Y_1, Y_2, Y_3)?