

# ST3131 Regression Analysis

SEMESTER II QUIZ 2024-2025

May 2025

TIME ALLOWED: 2 HOURS

---

## INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FIVE (5)** questions and comprises **FOUR (4)** printed pages.
2. Answer all questions. The marks for each question are indicated at the beginning of each question.
3. Answer each question beginning on a **FRESH** page of the answer book.
4. This is a **CLOSE BOOK** exam.
5. Candidates may use calculators. However, they should write down systematically the steps in the workings.

1. Suppose  $r = 1$ . Which of the following is true?

- (i)  $\text{cov}(x, y) = \text{SD}(x)\text{SD}(y)$   
(ii) The points  $(x_i, y_i), 1 \leq i \leq n$  lie on a straight line with slope 1
- |                       |                          |
|-----------------------|--------------------------|
| (A) Only (I)          | (B) Only (II)            |
| (C) Both (I) and (II) | (D) Neither (I) nor (II) |

2. Refer to slide B1. For  $h = 64$ , which of the following is closest to  $m_h$ ?  
 (A) 66                      (B) 67                      (C) 68                      (D) 69                      (E) 70

## Quiz 3

(4 marks)

1. A census (complete study of a population) is conducted on individuals between 20 and 50 years old in a country. The scatter diagram of height vs age is like an ellipse, and the regression line of height on age has a slope of -0.3 cm per year. Which of the following can be concluded from the given information?

- (i) Focus on the 30 year-olds in the country. 10 years later, their mean height will be about 3 cm less than now.
- (ii) Associated with an increase of 10 years, there is a decrease in height by about 3 cm.
- (A) Only (I)
- (B) Only (II)
- (C) Both (I) and (II)
- (D) Neither (I) nor (II)

2. For the Pearson data set, the father's heights ( $y$ ) have an SD of 2.8 inches, to one decimal place. Consider points with  $x$  values between 69.5 and 70.5 inches in the diagram on slide B2. The SD of the  $y$  values of these points is
- (A)  $> 2.8$                                       (B)  $\approx 2.8$                                       (C)  $< 2.8$

**Quiz 4**

(4 marks)

1. In a census on all married couples in a big city, the scatter diagram of  $y$  (husband's IQ) vs  $x$  (wife's IQ) is like an ellipse, and  $\bar{x} = \bar{y} = 100$ ,  $s_x = s_y = 10$ ,  $r = 0.6$ . For women with IQ 90, roughly, their husbands' mean IQ is
- (A) less than 90                                      (B) equal to 90  
 (C) between 90 and 94                                      (D) equal to 94  
 (E) more than 94
2.  $\hat{y} = mx + c$  be the predicted values from the regression line (i.e.,  $m = r \frac{s_y}{s_x}$ ,  $c = \bar{y} - m\bar{x}$ ), and  $e = y - \hat{y}$  be the residuals. Which of the following is true?
- (i) If every  $x_i, y_i$  is positive, then  $m > 0$ .  
 (ii) Given the values of  $e_1, e_2, \dots, e_{n-1}$ , we can determine the value of  $e_n$
- (A) Only (I)                                      (B) Only (II)  
 (C) Both (I) and (II)                                      (D) Neither (I) nor (II)

**Quiz 5**

(4 marks)

1. Let  $\hat{y} = mx + c$  be the predicted values based on the regression line with gradient  $m$  and  $y$ -intercept  $c$ . It is -- true that  $\widehat{\bar{y}} = \bar{\hat{y}}$ .

(A) always                      (B) sometimes                      (C) never

2. In certain research circles,  $r^2 \geq 0.25$  is considered to indicate a "large effect". A hypothetical author states "About 35% of the variation in  $y$  is explained by the regression on  $x$ ." The ratio of the regression RMSE to  $s_y$  is closest to

(A) 0.4                      (B) 0.5                      (C) 0.6                      (D) 0.7                      (E) 0.8

**Quiz 6**

(4 marks)

1. Given data  $x, y$  of length  $n > 1$  and  $s_x > 0$ , let the regression line of  $y$  on  $x$  have slope  $m$  and  $y$ -intercept  $c$ . Let  $\hat{y}$  be the predicted values using the regression line.  $\text{cov}(\hat{y}, x)$  is ----- equal to  $\text{cov}(y, x)$ .

(A) always                      (B) sometimes                      (C) never

2. Let  $x, y \in \mathbb{R}^n$ , where  $n > 1$ . Which of the following is true?

(i) If  $\text{cov}(x, y) = 0$ , then  $x \cdot y = 0$ .

(ii) If  $x \cdot y = 0$ , then  $\text{cov}(x, y) = 0$ .

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

**Quiz 7**

(4 marks)

1. Suppose  $x, y \in \mathbb{R}^n$  where  $n > 1$ , with means  $\bar{x}, \bar{y}$  and  $|r| < 1$ . Let  $m$  and  $c$  be the slope and  $y$ -intercept of the regression line of  $y$  on  $x$ , and  $\hat{y}$  be the corresponding predicted values. Which of the following is true?

- (i)  $y - \bar{y} = m(x - \bar{x})$
- (ii)  $\hat{y} - \bar{y} = m(x - \bar{x})$
- (A) Only (I)
- (B) Only (II)
- (C) Both (I) and (II)
- (D) Neither (I) nor (II)

2. Refer to slide 2.A3. Which of the following is true?

- (i)  $\beta_1$  is a random variable.
- (ii)  $\epsilon_1, \dots, \epsilon_n$  are random variables.
- (A) Only (I)
- (B) Only (II)
- (C) Both (I) and (II)
- (D) Neither (I) nor (II)

**Quiz 8**

(4 marks)

1. Refer to the display on the top of 2.A4:  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, 1 \leq i \leq n$ . Suppose  $x_1 = x_2$ . Which of the following is true?

- (i)  $Y_1 = Y_2$
- (ii)  $Y_1, Y_2$  have the same distribution
- (A) Only (I)
- (B) Only (II)
- (C) Both (I) and (II)
- (D) Neither (I) nor (II)



**Quiz 10**

(4 marks)

1. Let  $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ ,  $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$  and  $X = [1 \ x]$ ,  $E(Y)$  is

(i)  $X\beta$

(ii)  $\beta^\top X^\top$

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

2. Which of the following is true?

(i) Always  $\sum_{i=1}^n \epsilon_i = 0$

(ii)  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are independent RVs if and only if  $s_x^2 = \overline{x^2}$ .

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

**Quiz 11**

(4 marks)

Using data generated from the model specified in 2.A3 and 2.A4, a statistician (who does not know the values of  $\beta_0, \beta_1, \sigma$ ) estimates  $\beta_1$  as 0.42

1. Let  $\hat{\beta}_1$  be the LS estimator of  $\beta_1$ .  $\Pr(\hat{\beta}_1 \neq 0.42) =$   
 (A) 0                      (B) 0.42                      (C) 1                      (D) None of above
  
2. Which of the following is always true?
  - (i) Since 0.42 is an unbiased estimate,  $\beta_1 = 0.42$  (to two decimal places).
  - (ii) 0.42 is contained in the  $(1 - \alpha)$ -CI for  $\beta_1$ , where  $0 < \alpha < 1$ .
 (A) Only (I)                      (B) Only (II)  
 (C) Both (I) and (II)                      (D) Neither (I) nor (II)

**Quiz 12**

(4 marks)

1. Let  $A$  and  $b$  be  $n \times k$  and  $k \times 1$  matrices, and  $z = Ab$ . Which of the following is true?
  - (i) For  $i = 1, \dots, n$ ,  $z_i$  is the dot product of (row  $i$  of  $A$ ) and  $b$
  - (ii)  $z$  is a linear combination of columns  $1, \dots, m$  of  $A$ , with respective coefficients  $b_1, \dots, b_m$ .
 (A) Only (I)                      (B) Only (II)  
 (C) Both (I) and (II)                      (D) Neither (I) nor (II)

2. The rank of matrix  $\begin{bmatrix} -1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 4 \end{bmatrix}$  is  
 (A) 0                      (B) 1                      (C) 2                      (D) None of above



**Quiz 13**

(4 marks)

Let  $n > p + 1$  and  $X$  be the  $n \times (p + 1)$  matrix  $\begin{bmatrix} 1 & x_1 & \cdots & x_p \end{bmatrix}$  of rank  $p + 1$ . Let  $y$  be  $n \times 1$ . Let the LS hyperplane coefficients be  $b_{\text{ls}}$ , and  $e$  be the corresponding residuals.

1. The last entry of  $X^\top y$  is equal to

(i)  $y^\top x_p$

(ii)  $yx_p^\top$

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

2. Which of the following is true?

(i)  $X^\top X b_{\text{ls}} = X^\top y$

(ii)  $e^\top X$  contains only zero's

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

## Quiz 14

(4 marks)

Assume:

- $n > p + 1$ ,  $y$  is  $n \times 1$ .
- $X = \begin{bmatrix} 1 & x_1 & \cdots & x_p \end{bmatrix}$  is  $n \times (p + 1)$  of rank  $p + 1$
- $b_{ls}$ : LS hyperplane coefficients,  $e$  corresponding residuals.

1. Which of the following is true?

- (i)  $e$  and  $Xb_{ls}$  are orthogonal.
- (ii) If  $b \neq b_{ls}$ , it is possible that  $e$  and  $Xb$  are not orthogonal

- (A) Only (I)  
 (B) Only (II)  
 (C) Both (I) and (II)  
 (D) Neither (I) nor (II)

2. Let  $H = X(X^\top X)^{-1}X^\top$ . Which of the following is true?

- (i) The product of  $H$  and  $I - H$  is the zero matrix.
- (ii)  $(I - H)^2 = I - H$

- (A) Only (I)  
 (B) Only (II)  
 (C) Both (I) and (II)  
 (D) Neither (I) nor (II)

**Quiz 15**

(4 marks)

Assume:

- $n > p + 1$ ,  $y$  is  $n \times 1$ .
- $X = [\mathbf{1} \ x_1 \ \cdots \ x_p]$  is  $n \times (p + 1)$  of rank  $p + 1$ .

1. Which of the following is true?

(i)  $\text{rank } X^\top X = p + 1$

(ii)  $\text{rank } XX^\top = p + 1$

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

2. Let  $H = X(X^\top X)^{-1}X^\top$ . Which of the following is true?

(i)  $H\mathbf{1} = \mathbf{1}$

(ii) The column space of  $H$  consists of vectors of  $\mathbb{R}^{p+1}$ .

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

**Quiz 16**

(4 marks)

1. In the Hald dataset, which predictor has the largest absolute correlation with  $y$ 

(A) x1

(B) x2

(C) x3

(D) x4

(4 marks)

- 12

**Quiz 18**

(4 marks)

The questions are based on this extract from an output of `summary()` in R.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	36.8713	7.7090	4.783	5.02e-05 ***
x	0.4757	0.1131	4.204	0.000243 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.899 on 28 degrees of freedom

Multiple R-squared: 0.3869, Adjusted R-squared: 0.3651

F-statistic: 17.67 on 1 and 28 DF, p-value: 0.0002427

- Assume the "Multiple R-squared" is exactly 0.3869. Let  $r$  be the correlation between the variables.
  - $r = \sqrt{0.3869}$
  - $r = -\sqrt{0.3869}$
  - Either (A) or (B) is true, but there is insufficient information to decide.
  - None of the above.
- Let  $P_1$  be the p-value from the t-test of  $H_0: \beta_1 = 0$ .  
 Let  $P_2$  be the p-value from the  $F$ -test referred to by the last line.
 

(A) $P_1 < P_2$	(B) $P_1 = P_2$	(C) $P_1 > P_2$
-----------------	-----------------	-----------------

**Quiz 19**

(4 marks)

1. A data scientist wants to regress  $y$  on two categorical variables, one with 2 categories, and the other with 3 categories. The number of columns in the design matrix is
 

(A) 6
(B) 5
(C) 4
(D) 3
  
2. Let  $SSR_1$  and  $SSE_1$  be the regression and error SS from fitting a multiple regression. Let  $SSR_0$  and  $SSE_0$  be the regression and error SS from fitting a submodel, where some predictors are left out. Which of the following is true?
 

(i)  $SSE_1 \leq SSE_0$

(ii)  $SSR_1 - SSR_0 = SSE_0 - SSE_1$

(A) Only (I)
(B) Only (II)

(C) Both (I) and (II)
(D) Neither (I) nor (II)

**Quiz 20**

(4 marks)

1. Which of the following is not a column of the design matrix in the regression object `mod1` in `Lec5.R`?
 

(A) (1,1,1,1, 0,0,0,0, 0,0,0,0,0)  
(B) (0,0,0,0, 1,1,1,1, 0,0,0,0,0)  
(C) (0,0,0,0, 0,0,0,0, 1,1,1,1,1)

2. Consider the multiple regression model in slide 2 of 5\_misc.pdf, where  $p > 1$ . Let `mod` be created by `lm()` that regresses  $y$  on  $x_1, \dots, x_p$ . For  $j = 1, \dots, p$ , the  $P$  value against  $H_0 : \beta_j = 0$  can be read off the row for  $x_j$  in
- |                               |                          |
|-------------------------------|--------------------------|
| (i) <code>summary(mod)</code> |                          |
| (ii) <code>anova(mod)</code>  |                          |
| (A) Only (I)                  | (B) Only (II)            |
| (C) Both (I) and (II)         | (D) Neither (I) nor (II) |

## Quiz 21

(4 marks)

1. Assume the usual multiple regression model. A data analyst wants to find out if  $x_2$  should be used for predicting  $y$ . Which of the following is true?
  - (i) The null hypothesis says the estimate of  $\beta_2$  is 0.
  - (ii) The null hypothesis says the estimator of  $\beta_2$  is 0.

(A) Only (I)

(C) Both (I) and (II)

(B) Only (II)

(D) Neither (I) nor (II)
2. Let  $n > p + 1$ , and the  $n \times (p + 1)$  matrix  $X = [ \begin{array}{cccc} 1 & x_1 & \cdots & x_p \end{array} ]$  have rank  $p + 1$ . Which of the following is true about the diagonal entries of  $H = X(X^\top X)^{-1}X^\top$ ?
  - (A)  $h_{ii} \leq 1$  for  $i = 1, \dots, n$
  - (B)  $h_{ii} > 1$  for  $i = 1, \dots, n$
  - (C) None of the above.

**Quiz 22**

(4 marks)

1. Let  $j \in \{1, \dots, p\}$  be the LS estimate of  $\beta_j$ ,  $\hat{\sigma}^2$  be the unbiased estimate of  $\sigma^2$ , and  $\eta_j$  be the  $(j+1)$ -diagonal entry of  $(X^\top X)^{-1}$  observed test statistic for testing  $H_0 : \beta_j = 0$  is

$$t = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{\eta_j}}$$

The P-value is

- (i)  $Pr(|t_{n-p-1}| \geq |t|)$
- (ii)  $Pr(F_{1,n-p-1}^2 \geq t^2)$
- (A) Only (I)
- (B) Only (II)
- (C) Both (I) and (II)
- (D) Neither (I) nor (II)

2. Which of the following is true?

- (i)  $E(Y) = X\beta$
- (ii)  $E(\hat{Y}) = X\beta$
- (A) Only (I)
- (B) Only (II)
- (C) Both (I) and (II)
- (D) Neither (I) nor (II)



**Quiz 23**

(4 marks)

1. Let  $X$  be an  $n \times m$  matrix, with  $n > m$  and  $\text{rank } X = m$ . Which of the following is true?

(i)  $H_1 X = X$ , where  $H_1 = X (X^\top X)^{-1} X^\top$

(ii)  $H_2 X^\top = X^\top$ , where  $H_2 = X^\top (X X^\top)^{-1} X$

(A) Only (I)

(B) Only (II)

(C) Both (I) and (II)

(D) Neither (I) nor (II)

2. For the model in 6.B1, let  $Y_i = \frac{1}{J_i} \sum_{j=1}^{J_i} Y_{ij}$ ,  $\text{var}(Y_{i\cdot}) =$

(A)  $\frac{\sigma^2}{J_i}$       (B)  $\frac{\sigma^2}{J_i-1}$       (C)  $\frac{\sigma^2}{J_i-2}$       (D)  $\frac{\sigma^2}{J_i-\mathcal{I}}$       (E) None of the above

**END OF PAPER**