Department of Computational Mathematics, Faculty of Information Technology
University of Moratuwa

Honours Degree of Bachelor of science in Artificial Intelligence.

Batch 21 - Level 2 (Semester II)

CM 2111: Statistical Inference

Tutorial 06

01. A random sample of 11 statistics students produced the following data, given the third exam score out of 80, and the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Third exam score	65	67	71	71	66	75	67	70	71	69	69
Final exam score	175	133	185	163	126	198	153	163	159	151	159

a. What is an explanatory variable, x and response variable, y?
b. Make scatter plot to see the relationship between the variable x and y.
c. Find the least square regression line.
d. Predict the final exam score, if the third exam score is 76.

02. The equation of a least-squares regression line is $y = 10 + 5x$.
a. What is the slope and intercept of the regression line?
b. If x increases by one unit, what is the corresponding increase in y?
c. What is the value of y, for $x = 5$?

03. Nitrogen balance studies are used to determine protein requirements for people. Each subject is fed three different controlled diets during three separate experimental periods. The three diets are similar with regard to all nutrients except protein. Nitrogen balance is the difference between the amount of nitrogen consumed and the amount lost in feces and urine and by other means. Since virtually all of the nitrogen in a diet comes from protein, nitrogen balance is an indicator of the amount of protein retained by the body. The protein requirement for an individual is the intake corresponding to a balance of zero. Linear regression is used to model the relationship between nitrogen balance, measured in milligrams of nitrogen per kilogram of body weight per day (mg/kg/d), and protein intake, measured in grams of protein per kilogram of body weight per day (g/kg/d).

Protein intake	0.543	0.797	1.030
Nitrogen balance	−23.4	17.8	67.3

Here is the summary of the data set.

Variable	Mean	Standard Deviation	Correlation Coefficient	Sample size
Protein Intake	0.79000	0.24357545	0.99698478	3
Nitrogen Balance	20.56667	45.4132506		

a. What is an explanatory variable, x and response variable, y?
b. Assuming the linear regression model $y = \beta_0+\beta_1 x+\epsilon$, estimate the values for β_0 and β_1 and then write down the equation of least square regression line.
c. Predict the nitrogen balance, if the protein intake is 0.543.
d. Compute residuals for intakes.
e. Calculate the estimate of the standard deviation about the line.
f. Find the standard error of the slope b_1 of the least-squares regression line.
g. Test the null hypothesis $\beta_1 = 0$ against the alternative hypothesis $\beta_1 \neq 0$ at 0.05 level of significance.
h. Find a 95% confidence interval for the slope β_1.

04. A researcher wants to estimate the relationship between the daily temperature (x) and the amount of electricity (y) used by customers. The following data were collected.

x	85	90	76	91	84	94	88	85	97	86	82	78	77	83
y	22.5	23.7	20.3	23.4	24.2	23.5	22.9	22.4	26.1	23.1	22.5	20.9	21	22.6

a. Plot the data.
b. Assuming the linear regression model $y = \beta_0+\beta_1 x+\epsilon$, estimate the values for β_0 and β_1 and then write down the equation of least square regression line.
c. Predict the amount of electricity used by customers, if the daily temperature is 95 and 87.

05. The following ANOVA table, based on information with employee performance ratings as the dependent variable and years of experience, education level, training hours and department as independent variables, has a few missing values.

Source of Variation	**DF**	**Sum of Square**	**Mean Square**	**F-value**
Regression	1	b	19.2813	e
Error	a	89.3677	d	
Total	12	c		

Find the missing values a, b, c, d and e, and complete the ANOVA table.

06. An insurance company wants to investigate the relationship between the income of person (in thousands of dollars) and the amount of life insurance (in thousands of dollars). The research department at the company collected information on 8 persons.

Annual Income	62	78	41	53	85	34	45	65
Amount of Life insurance	250	300	100	150	500	75	125	265

a. Write down the independent variable and the dependent variable.
b. Make the scatter plot to see the relationship between the annual income and amount of life insurance.
c. Assuming the linear regression model $y = \beta_0+\beta_1 x+\epsilon$, estimate the values for β_0 and β_1 and then write down the equation of least square regression line.
d. Construct the analysis of variance table. (ANOVA table)
e. Test the null hypothesis $\beta_1 = 0$ against the alternative hypothesis $\beta_1 \neq 0$ at 0.05 level of significance (using ANOVA table).
f. Calculate the coefficient of determination and interpret it.

07. A large mid-western bank is planning on introducing a new word processing system to its secretarial staff. To learn about the amount of training that is needed to effectively implement the new system, the bank chose eight employees of roughly equal skill. These workers were trained for different amounts of time and were then individually out to work on a given project. The following data indicate the training times and the resulting times (both in hours) that it took each worker to complete the project.

Training Time	**Time to Complete Project**
22	18.4
18	19.2
30	14.5
16	19
25	16.6
20	17.7
10	24.4
14	21

a. Construct a scatter diagram for the data.
b. Assuming the linear regression model $y = \beta_0+\beta_1 x+\epsilon$, estimate the values for β_0 and β_1.
c. What is the estimated regression line?
d. Construct the ANOVA table to test the significance of the regression parameters.
e. Test the null hypothesis $\beta_1 = 0$ against the alternative hypothesis $\beta_1 \neq 0$ at 0.05 level of significance (using ANOVA table).
f. Calculate the coefficient of determination and interpret it.
g. Predict the amount of time it would take a worker who receives 28 hours of training to complete the project.

08. The fitted regression equation for a multiple regression is

$$\hat{y} = -1.4 + 2.6x_1 - 2.3x_2$$

a. If $x_1 = 4$ and $x_2 = 2$, what is the predicted value of y?
b. For the answer to part (a) to be valid, is it necessary that the values $x_1 = 4$ and $x_2 = 2$ correspond to a case in the data set? Explain why or why not.
c. If you hold x_2 at a fixed value, what is the effect of an increase of two units in x_1 on the predicted value of y?

09.

a. A retail chain wants to optimize its inventory management to maximize monthly sales. The company has historical data on monthly sales (y), the number of products in stock (x_1), and the advertising expenditure (x_2). The goal is to build a model that can predict monthly sales based on these variables, helping the company make informed decisions about inventory levels and - advertising budgets.

y	9	10	13	14	16
x_1	1	3	4	6	7
x_2	10	14	15	18	20

i. Find the coefficient of the regression using the matrix form.
ii. Write down the multiple linear regression model.

b. Consider the following data set and find the ANOVA table for linear regression.
(Hint: The Regression equation is equal to y= $1.5543+0.3722x_1+0.0442x_2$ and the total sum of square is 17.2)

y	1	3	5	6	2
x_1	1	3	4	6	7
x_2	2	10	11	2	7