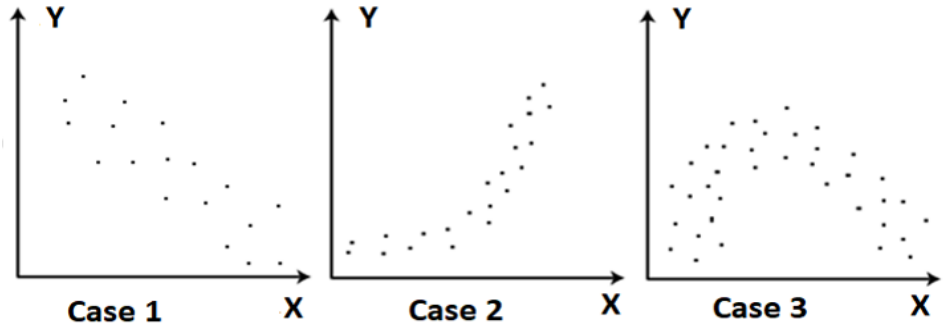
Tutorial Worksheet 4 - Regression Analysis

Objective Questions

1. Which of the following three cases depicts a 'non-monotonic relationship' between the two variables X and Y?



   (a) Case 1 (Plot in the left) (b) Case 2 (Plot in the centre) (c) Case 3 (Plot in the right) (d) None of the plots depicts 'non-monotonic relationship'

2. Look into the following three statements carefully:

   (I) Karl Pearson's Correlation Analysis method can be used to find correlation coefficient between two numerical attributes.

   (II) Charles Spearman's Correlation Analysis method can be used to find correlation coefficient between two ordinal attributes

   (III) Chi-square Coefficient of Correlation Analysis method can be used to find correlation coefficient between two nominal attributes

   Which of the following options is TRUE?

   (a) Only Statement I is correct, other two statements are incorrect

   (b) Only Statement II is correct, other two statements are incorrect

   (c) Only Statement III is correct, other two statements are incorrect

   (d) All three statements, I, II, and III are correct

3. In order to compute the Charles Spearman's Coefficient of Correlation between two variables, a rank is assigned to each data. Further, the ranks are modified/corrected if two data are of same value. The observations of such two variables X and Y are given in below table.

X	5	10	15	20	25
Y	200	300	300	300	500

   What will be the modified (final) rank of the observations of variable Y = 200, 300, 300, 300, 500?

   (a) 5, 4, 3, 2, 1 (b) 3, 2, 2, 2, 1 (c) 5, 3, 3, 3, 1 (d) 5, 4.5, 3.5, 2.5, 1.

4. In regression analysis, the variable that is being predicted is called as?

   (a) Response (b) Regressor (c) Independent variable (d) Dependent variable

5. In the least square method of regression analysis, what is the relationship between the sum of squares of the errors (SSE), total corrected sum of squares (SST) and the coefficient of determination (R^2)?

(a) $R^2 = 1 - \frac{SSE}{SST}$ (b) $R^2 = 1 + \frac{SSE}{SST}$ (c) $R^2 = 1 - \frac{SST}{SSE}$ (d) $R^2 = 1 + \frac{SST}{SSE}$

6. If the coefficient of determination is a positive value, then the regression equation?

   (a) must have a positive slope (b) must have a negative slope (c) could have either a positive or a negative slope (d) must have a positive y intercept

7. What is (are) the required assumption(s) for the auto-regression analysis?

   (a) The time series under consideration is non-stationary.

   (b) The time series under consideration is non-uniform.

   (c) The time series under consideration is stationary, but not uniform.

   (d) The time series under consideration is both stationary and uniform.

8. For a given dataset, to compute the relationship between the variables x and y, following two regression models are obtained.

   (I) Model 1: $\mathrm{Y} = 2\mathrm{X}_2 + \mathrm{X}_1$, with R^2 score $= 0.68$

   (II) Model 2: $\mathrm{Y} = 3\mathrm{X}_3 + 2\mathrm{X}_2 + \mathrm{X}_1$ with R^2 score $= 0.87$

   Which model is more acceptable?

   (a) a) Model 1 (b) Model 2 (c) Both models are equally acceptable (d) None of the model is acceptable

9. In a regression analysis if the coefficient of determination $R^2 = 1$, then sum of squares of the errors (SSE) must be equal to?

   (a) 1 (b) 0 (c) Any positive value (d) Infinity

10. A study was conducted to investigate the relationship between a student's pocket-money in AED per year and his/her family income in AED per year. Data of the same were collected for 100 students, and a simple linear regression line of the form $Y = \alpha + \beta X$ was fitted. Now suppose that the data of both the family income and pocket money is converted from AED to USD. The impact of this conversion on the regression line is

    (a) The sign of the slope will change, but the magnitude of the slope will remain unchanged

    (b) The magnitude of the slope will change, but the sign of the slope will remain unchanged

    (c) Both the sign and magnitude of the slope will change

    (d) None of the sign and magnitude of the slope will change

Subjective Questions

Problem 1.

The ranks for 10 students on the foundation year examination (A) and L3 examination (B) are given in the table below. What is Spearman's rank correlation coefficient between A and B?

Rank in A	5	3	10	8	2	6	9	4	1	7
Rank in B	6	2	10	1	4.5	4.5	8	3	9	7

Problem 2.

In order to find out the correlation between an independent variable X and a dependent variable Y, following information is available.

$$\sum \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) = 498, \sum \left(X_i - \bar{X}\right)^2 = 338, \sum \left(Y_i - \bar{Y}\right)^2 = 1212$$

What is the value of Karl Pearson's coefficient of Correlation between X and Y ?

Problem 3.

The sigmoid function takes the form $f(x) = \frac{1}{1+e^{-x}}$. What is the first order differentiation of $f(x)$ w.r.t to $\mathbf{x}$?

Problem 4.

A simple linear regression model of the form $Y = a + bX$ is used to compute the relationship between the variables X and Y. Suppose there are n sample points, $(x_i, y_i), i = 1, 2, \ldots, n$, and $\bar{x}$ and $\bar{y}$ are their corresponding means. The value of linear regression model coefficient b is given by?

Problem 5.

What are the maximum and minimum values of the function $p(x) = \frac{1}{1+e^{-\frac{-(x-\mu)}{s}}}$, for $-\infty < x < \infty$, and $\mathbf{s}, \boldsymbol{\mu}$ are constants?

Problem 6.

Studies have shown that people who suffer sudden cardiac arrest have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? The accompanying data give y = survival rate (percent) and $x = 5$ mean call-to shock time (minutes) for a cardiac rehabilitation center (in which cardiac arrests occurred while victims were hospitalized and so the call-to-shock time tended to be short) and for four communities of different sizes:

Mean call-to-shock-time, x :	2	6	7	9	12
Survival rate, y :	90	45	30	5	2

(a) Construct a scatterplot for these data. How would you describe the relationship between mean call-to shock time and survival rate?

(b) Find the equation of the least-squares line.

(c) Use the least-squares line to predict survival rate for a community with a mean call-to-shock time of 10 minutes.

Problem 7.

Let x be the size of a house (in square feet) and y be the amount of natural gas used (therms) during a specified period. Suppose that for a particular community, x and y are related according to the simple linear regression model with β = slope of population regression line = 0.017, $\alpha = y$ intercept of population regression line = -5.0

Houses in this community range in size from 1000 to 3000 square feet.

(a) What is the equation of the population regression line?

(b) Graph the population regression line by first finding the point on the line corresponding to $x = 1000$ and then the point corresponding to $x = 2000$, and drawing a line through these points.

(c) What is the mean value of gas usage for houses with 2100 sq. ft. of space?

(d) What is the average change in usage associated with a 1 sq. ft. increase in size?

(e) What is the average change in usage associated with a 100 sq. ft. increase in size?

Problem 8.

The data in the accompanying table is from the paper "Six-Minute Walk Test in Children and Adolescents" (The Journal of Pediatrics [2007]: 395-399). Two hundred and eighty boys completed a test that measures the distance that the subject can walk on a flat, hard surface in 6 minutes. For each age group shown in the table, the median distance walked by the boys in that age group is also given.

Age Group	Representative Age (Mid point)	Median Distance (Meters)
$3-5$	4	$544 \cdot 3$
$6-8$	7	584.0
$9-11$	10	$667 \cdot 3$
$12-15$	13.5	701.1
$16-18$	17	727.6

(a) With x = representative age and y = median distance walked in 6 minutes, construct a scatterplot. Does the pattern in the scatterplot look linear?

(b) Find the equation of the least-squares regression line that describes the relationship between median distance walked in 6 minutes and representative age.

(c) Compute the five residuals and construct a residual plot. Are there any unusual features in the plot?

Problem 9.

A simple linear regression model was used to describe the relationship between y = hardness of molded plastic and x = amount of time elapsed since the end of the molding process. Summary quantities included $n = 15$, SSResid = 1235.470, and SSTo = 25,321.368.

(a) Calculate a point estimate of σ. On how many degrees of freedom is the estimate based?

(b) What percentage of observed variation in hardness can be explained by the simple linear regression model relationship between hardness and elapsed time?

Problem 10.

An experiment to study the relationship between x = time spent exercising (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \quad \sum x = 50, \quad \sum y = 16705,$$
$$\sum x^2 = 150, \quad \sum y^2 = 14194231, \quad \sum xy = 44194$$

(a) Estimate the slope and y intercept of the population regression line.

(b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

(c) Compute a 99% confidence interval for the average change in oxygen consumption associated with a 1-minute increase in exercise time.

Problem 11.

A simple linear regression model was used to describe the relationship between sales revenue y (in thousands of dollars) and advertising expenditure x (also in thousands of dollars) for fast-food outlets during a 3-month period. A sample of 15 outlets yielded the accompanying summary quantities.

$\sum x = 14.10, \quad \sum y = 1438.50, \quad \sum x^2 = 13.92, \quad \sum y^2 = 140354$
$\sum xy = 1387.20, \quad \sum (y - \bar{y})^2 = 2401.85, \quad \sum (y - \hat{y})^2 = 561.46$

(a) What proportion of observed variation in sales revenue can be attributed to the linear relationship between revenue and advertising expenditure?

(b) Calculate s_e and s_b.

(c) Obtain a 90% confidence interval for β, the average change in revenue associated with a \$1000 (that is, 1-unit) increase in advertising expenditure.

(d) Test the hypothesis $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ using a significance level of 0.05. What does your conclusion say about the nature of the relationship between x and y?

Problem 12.

The article "Effect of Temperature on the pH of Skim Milk" (Journal of Dairy Research [1988]: 277- 280) reported on a study involving x = temperature (°C) under specified experimental conditions and y = milk pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x :	4	4	24	24	25	38	38	40	45	50	55	56	60	67	70	78
y :	6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \sum y = 104.54 \sum x^2 = 36,056 \sum y^2 = 683.4470 \sum xy = 4376.36$

(a) Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

(b) Obtain a 95% confidence interval for $\alpha + \beta(40)$ the mean milk pH when the milk temperature is 40°C.

(c) Obtain a 99% prediction interval for a single pH observation to be made when milk temperature = 40°C.

(d) Would you recommend using the data to calculate a 95% confidence interval for the mean pH when the temperature is 90°C ? Why or why not?