Unit III

Coefficient of Correlation

A coefficient of correlation is generally applied in statistics to calculate a relationship between two variables. The correlation shows a specific value of the degree of a linear relationship between the X and Y variables, say X and Y. There are various types of correlation coefficients. However, Pearson's correlation (also known as Pearson's R) is the correlation coefficient that is frequently used in linear regression.

Pearson's Coefficient Correlation

Karl Pearson's coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by **"r".**

Karl Pearson Correlation Coefficient Formula

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Alternative Formula (covariance formula)

$$\text{Cov}(X,Y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N} = \frac{\Sigma xy}{N}$$

Properties:

- Ranges from -1 to 1.
- r=1: Perfect positive correlation.
- r=−1: Perfect negative correlation.
- r=0: No correlation.

Interpretation:

- The closer the value of r is to 1 or -1, the stronger the correlation.
- The closer the value is to 0, the weaker the correlation.

Ranking Methods:

In some cases, variables may not be measured on a continuous scale, or the relationship may not be strictly linear. In such situations, ranking methods can be employed.

Unit III

Spearman's Rank Correlation Coefficient:

- Measures the strength and direction of monotonic association between two variables.
- Uses the ranks of the data rather than the actual values.

Kendall's Tau:

- Another non-parametric measure for the strength and direction of association.
- Focuses on concordant and discordant pairs of data points.

Key Points:

Assumptions of Pearson Correlation:

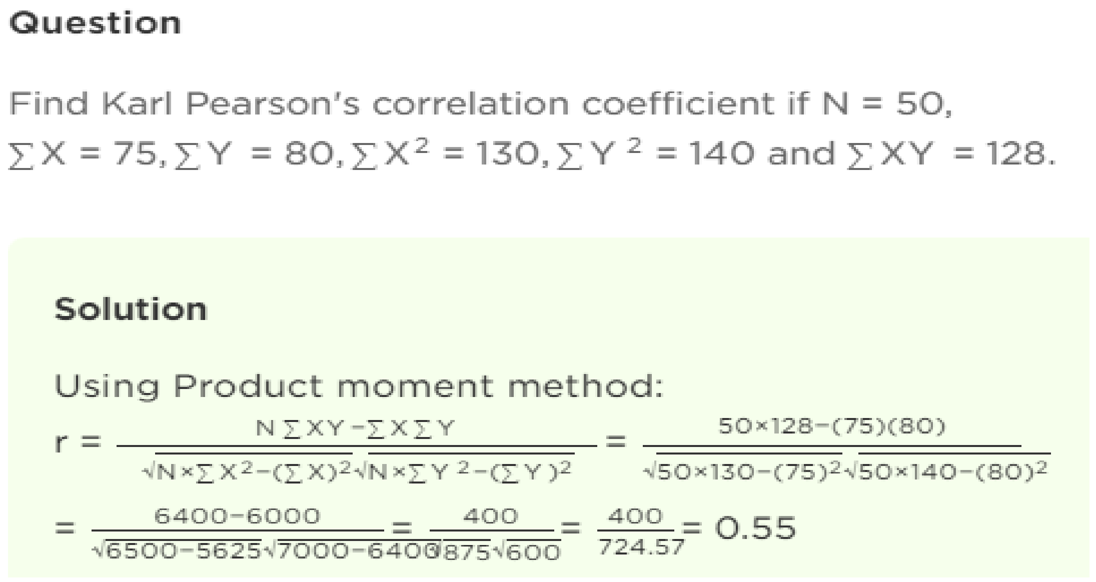
- Linearity: Assumes a linear relationship.
- Homoscedasticity: Assumes constant variance.
- Independence: Assumes independence of observations.

Advantages of Ranking Methods:

- Robust to outliers.
- Applicable to non-linear relationships.

Considerations:

- Correlation does not imply causation.
- Outliers can strongly influence correlation values.



Question

Find Karl Pearson's correlation coefficient if N = 50, $\sum X = 75, \sum Y = 80, \sum X^2 = 130, \sum Y^2 = 140$ and $\sum XY = 128$.

Solution

Using Product moment method:

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\times\sum X^2-(\sum X)^2}\sqrt{N\times\sum Y^2-(\sum Y)^2}} = \frac{50\times128-(75)(80)}{\sqrt{50\times130-(75)^2}\sqrt{50\times140-(80)^2}}$$

$$= \frac{6400-6000}{\sqrt{6500-5625}\sqrt{7000-6400}} = \frac{400}{\sqrt{875}\sqrt{600}} = \frac{400}{724.57} = 0.55$$

Unit III

Question: Consider a class of 10 students who have received scores in a mathematics competition. The scores are as follows:
85,92,78,90,88,95,80,87,94,89
Using the dense ranking method, rank the students based on their scores. If there are ties, assign the average rank to the tied scores.

Answer:
To solve this problem, first, we need to arrange the scores in descending order:
95,94,92,90,89,88,87,85,80,78
Now, assign ranks based on the **dense ranking method**:
1,2,3,4.5,4.5,6,7,8,9,10

In this case, there is a tie for the 4th and 5th positions, so we take the average of the ranks (4 and 5) and assign it to both scores. Therefore, the ranks for the tied scores are 4.5.

So, the final ranks based on the dense ranking method are:
1,2,3,4.5,4.5,6,7,8,9,10

Regression:

Definition:

Regression is a statistical technique used to model the relationship between a dependent variable (also called the response variable) and one or more independent variables (predictors or explanatory variables). The goal is to understand and quantify the relationship between variables, make predictions, or infer causal relationships.

Types of Regression:

Simple Linear Regression:

- Involves one dependent variable and one independent variable.
- The relationship is represented by a straight line.

Multiple Linear Regression:

- Involves one dependent variable and multiple independent variables.
- The relationship is represented by a hyperplane in a multidimensional space.

Polynomial Regression:

- Allows for a relationship between variables to be modeled as an nth-degree polynomial.
- Captures non-linear relationships between variables.

Logistic Regression:

- Used for binary classification problems.
- Models the probability of an event occurring.

Unit III

Regression Expressions:

Simple Linear Regression:

$Y=\beta_0+\beta_1 X+\varepsilon$

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept (the value of Y when X is 0).
- β_1 is the slope (the change in Y for a one-unit change in X).
- ε is the error term.

Multiple Linear Regression:

$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n+\varepsilon$

- $X_1, X_2+\ldots X_n$ are the independent variables.
- $\beta_0, \beta_1, \beta_2 \ldots\ldots \beta_n$ are the coefficients.

Lines of Regression:

Definition:

Lines of regression represent the best-fitting lines that minimize the sum of the squared differences between observed and predicted values in a regression analysis.

1. Regression Line (for Simple Linear Regression):

- Represents the best linear relationship between two variables.
- Minimizes the sum of squared differences between observed and predicted values.

2. Line of Best Fit (for Multiple Linear Regression):

- Involves fitting a hyperplane in multidimensional space.
- Minimizes the sum of squared differences between observed and predicted values.

3. Residuals:

- The vertical distances between data points and the regression line.
- Residuals help evaluate how well the model fits the data.

4. Coefficient of Determination (R^2):

- A measure of the proportion of the variance in the dependent variable that is predictable from the independent variables.

Unit III

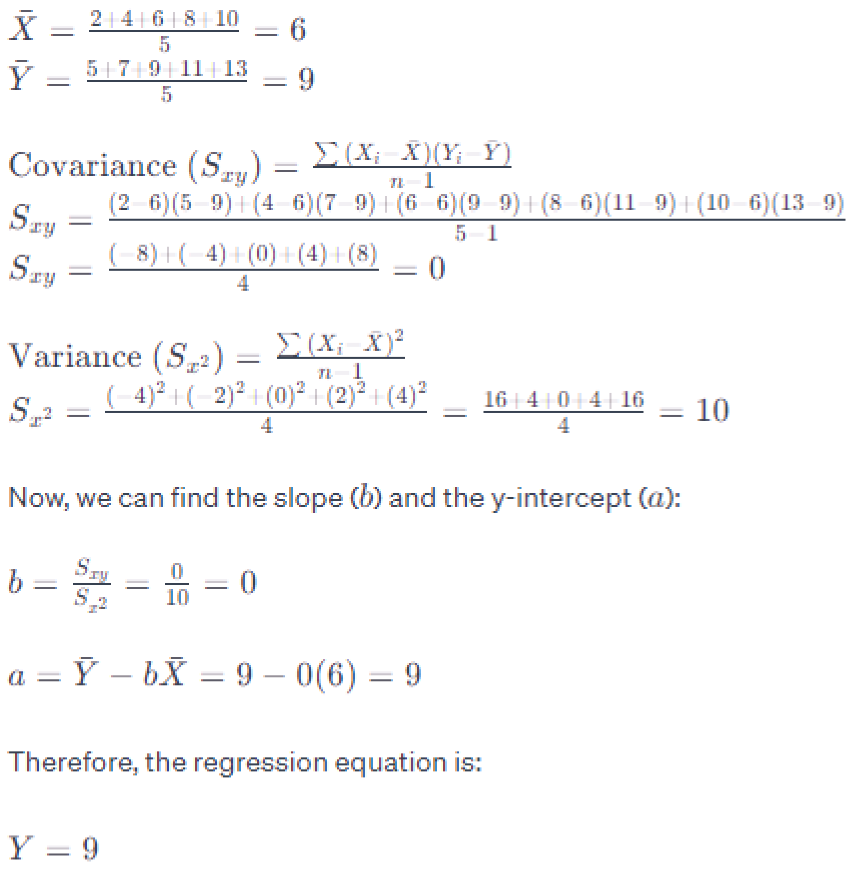
Question:

Suppose we have a dataset with the following values for two variables, X and Y:
X:2,4,6,8,10
Y:5,7,9,11,13

Calculate the regression equation for predicting Y based on X using the method of least squares. Provide the equation for the line of regression.

Answer:

To find the regression equation, we first need to calculate the mean of X and Y, as well as the covariance and variance.



$$\bar{X} = \frac{2+4+6+8+10}{5} = 6$$
$$\bar{Y} = \frac{5+7+9+11+13}{5} = 9$$

$$\text{Covariance } (S_{xy}) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
$$S_{xy} = \frac{(2-6)(5-9)+(4-6)(7-9)+(6-6)(9-9)+(8-6)(11-9)+(10-6)(13-9)}{5-1}$$
$$S_{xy} = \frac{(-8)+(-4)+(0)+(4)+(8)}{4} = 0$$

$$\text{Variance } (S_{x^2}) = \frac{\sum (X_i - \bar{X})^2}{n-1}$$
$$S_{x^2} = \frac{(-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2}{4} = \frac{16+4+0+4+16}{4} = 10$$

Now, we can find the slope (b) and the y-intercept (a):

$$b = \frac{S_{xy}}{S_{x^2}} = \frac{0}{10} = 0$$

$$a = \bar{Y} - b\bar{X} = 9 - 0(6) = 9$$

Therefore, the regression equation is:

$$Y = 9$$

In this case, the line of regression is a horizontal line with a y-intercept at 9, as the slope (b) is zero.

Interpolation and extrapolation :

Interpolation and extrapolation are statistical techniques used to estimate values between or beyond observed data points. In the context of interpolation, the binomial, Lagrange, and Newton methods are commonly employed. Let's delve into each method:

Binomial Interpolation:

1. Overview:

- The binomial interpolation method is primarily used for evenly spaced data points.
- It is based on the binomial coefficient formula.

2. Binomial Coefficient Formula:

- The binomial coefficient, often denoted as C(n, k), represents the number of ways to choose k elements from a set of n distinct elements.

3. Formula for Binomial Interpolation:

- The interpolation formula is given by:

$$P(x) = \sum_{k=0}^{n} \left[f(x_k) \cdot C(n,k) \cdot (x - x_0)^{n-k} \cdot (x_1 - x_0)^k \right]$$

Lagrange Interpolation:

1. Overview:

- Lagrange interpolation is a polynomial interpolation method.
- It constructs a polynomial that passes through all given data points.

2. Lagrange Polynomial:

- Given a set of n+1 data points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, the Lagrange polynomial is:

$$P(x) = \sum_{i=0}^{n} y_i \prod_{j=0, j \neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

3. Advantages:

- Simple and straightforward to implement.
- Works well for small to moderately sized datasets.

Newton's Divided Difference Interpolation:

1. Overview:

- Newton's method is based on divided differences and is also used for polynomial interpolation.

2. Divided Difference Formula:

- For a set of n+1 data points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, the divided difference is defined recursively as:

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$

3. Newton's Interpolation Formula:

- The interpolation polynomial is given by:

$$P(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \ldots$$

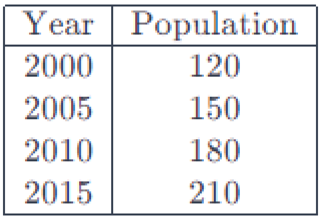
Extrapolation:

- While these methods are primarily designed for interpolation (estimating within the range of observed data), they can be extended for extrapolation (estimating beyond the observed range) with caution.
- Extrapolation involves using the same interpolation methods but extending the polynomial beyond the given data range.

Important Considerations:

- Interpolation and extrapolation are sensitive to the choice of method and may yield inaccurate results if not applied carefully.
- The choice of method depends on the nature of the data and the desired accuracy.

Question: Suppose you have the following data points representing the population (in thousands) of a city over the years:



Year	Population
2000	120
2005	150
2010	180
2015	210

1. Use the binomial method to estimate the population in the year 2020.
2. Use the Lagrange interpolation method to estimate the population in the year 2020.
3. Use the Newton interpolation method to estimate the population in the year 2020.

Assume that the trend is binomial, and you can use the degree of the binomial, Lagrange, and Newton methods accordingly.

Answer:

1. Binomial Method:

Assuming a binomial trend, the population can be estimated using the binomial formula:

$$P(t) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

where n is the number of years, k is the year of interest (2020 - 2000 = 20), p is the probability of success, and 1−p is the probability of failure.

Unit III

Let's assume p= 0.5 (for simplicity, considering an equal chance of increase or decrease in population).

$$P(20) = \binom{20}{10} \cdot 0.5^{10} \cdot 0.5^{10} \approx 0.176$$

The estimated population in 2020 using the binomial method is approximately 0.176×210,000=36,960 (in thousands).

2. Lagrange Interpolation Method:

The Lagrange interpolation polynomial for the given data points is:

$$P(t) = L_0(t) \cdot y_0 + L_1(t) \cdot y_1 + L_2(t) \cdot y_2 + L_3(t) \cdot y_3$$

where L_i (t) are Lagrange basis polynomials, and y_i are the corresponding function values.
Using the Lagrange basis polynomials, we can calculate P(2020):
P(2020)≈165.625
The estimated population in 2020 using the Lagrange interpolation method is approximately 165.625 (in thousands).

3. Newton Interpolation Method:

The Newton interpolation polynomial for the given data points is:
$P(t)=f[x_0]+(t-x_0)f[x_0 ,x_1]+(t-x_0)(t-x_1)f[x_0 ,x_1,x_2] +(t-x_0)(t-x_1)(t – x_2)f[x_0 ,x_1,x_2,x_3]$
where $f[x_i]$ are divided differences.
Using the divided differences, we can calculate P(2020):
P(2020)≈167.5
The estimated population in 2020 using the Newton interpolation method is approximately 167.5 (in thousands).