Lecture -10
Predicting missing data using regression modeling, interpolation

Vaibhav Chunekar

14 August 2023

Agenda

- Understand the Covariance and Correlation
- Prediction of Missing Value with Linear Regression
- Prediction of Missing Value with Multiple Linear Regression

Covariance

- In statistics and probability theory, covariance deals with the joint variability of two random variables: x and y.
- Generally, it is treated as a statistical tool used to define the relationship between two variables.
- Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.
- This is the property of a function of maintaining its form when the variables are linearly transformed.
- Covariance is measured in units, which are calculated by multiplying the units of the two variables

Types of Covariance

- Positive Covariance
- If the covariance for any two variables is positive, that means, **both the variables move in the same direction**. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.
- Negative Covariance
- If the covariance for any two variables is negative, that **means, both the variables move in the opposite direction.** It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa

Covariance formula

- Covariance formula is a statistical formula, used to evaluate the relationship between two variables.
- It is one of the statistical measurements to know the relationship between the variance between the two variables.
- Let us say X and Y are any two variables, whose relationship has to be calculated.
- Thus the covariance of these two variables is denoted by Cov(X,Y).

Covariance Formula

Population Covariance Formula

$$\mathrm{Cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$\mathrm{Cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - y)}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- x̄ = mean of x
- ȳ = mean of y
- N = number of data values.

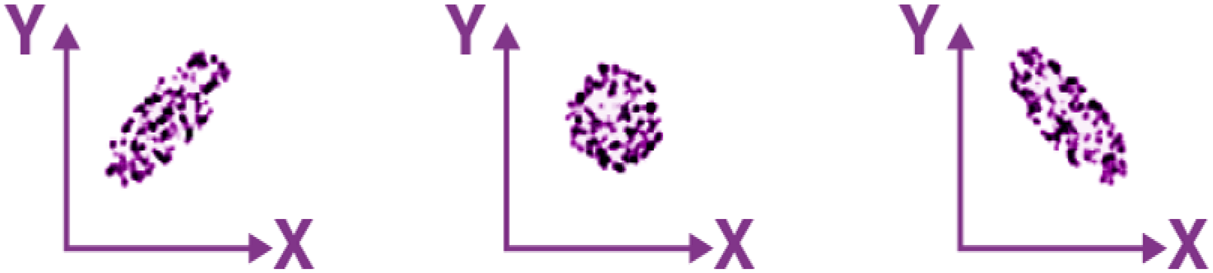
Example: Calculate the coefficient of covariance for the following data:

X	2	8	18	20	28	30
Y	5	12	18	23	45	50

Number of observations = 6
Mean of X = 17.67
Mean of Y = 25.5

$$\text{Cov}(X, Y) = (\tfrac{1}{6}) [(2 - 17.67)(5 - 25.5) + (8 - 17.67)(12 - 25.5) + (18 - 17.67)(18 - 25.5) + (20 - 17.67)(23 - 25.5) + (28 - 17.67)(45 - 25.5) + (30 - 17.67)(50 - 25.5)] = 157.83$$

Interpretations Covariance



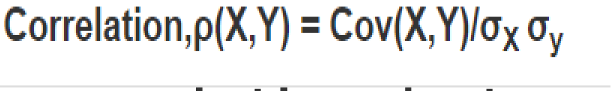
cov(X,Y)>0 cov(X,Y)≈0 cov(X,Y)<0

Interpretations Covariance

- If **cov(X, Y) is greater than zero**, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.
- If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.
- If cov(X, Y) is zero, then we can say that there is no relation between two variables.

Correlation Coefficient & Formula

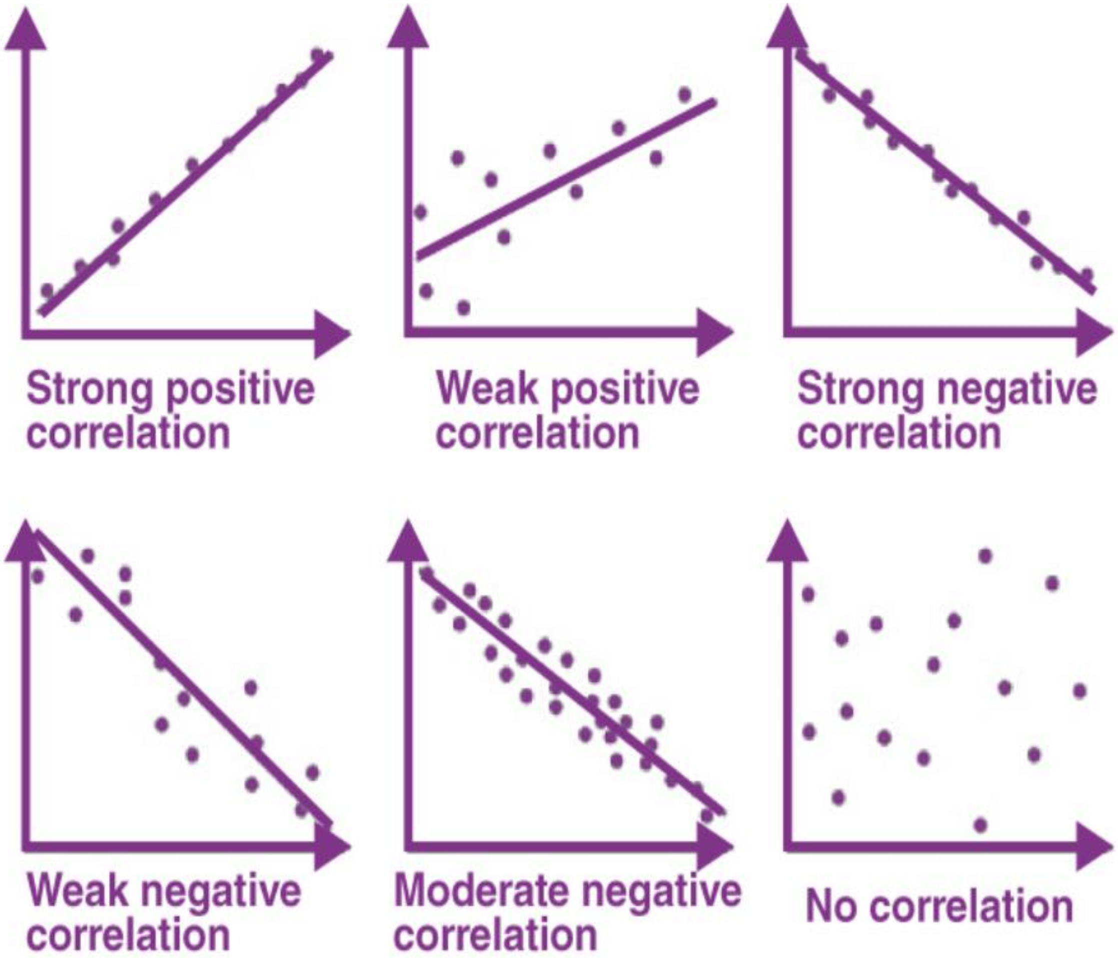
- Correlation estimates the depth of the relationship between variables. It is the estimated measure of covariance and is dimensionless. In other words, the correlation coefficient is a constant value always and does not have any units. The relationship between the correlation coefficient and covariance is given by:



Correlation,$\rho(X,Y) = \mathrm{Cov}(X,Y)/\sigma_X\,\sigma_y$

- ρ(X,Y) = correlation between the variables X and Y
- Cov(X,Y) = covariance between the variables X and Y
- σX = standard deviation of the X variable
- σY = standard deviation of the Y variable

Graphical representation of correlation among two variables



Strong positive correlation

Weak positive correlation

Strong negative correlation

Weak negative correlation

Moderate negative correlation

No correlation

Difference Between Covariance-Correlation

Covariance	**Correlation**
It is a measure to show the extent to which given two random variables change with respect to each other.	It is a measure used to describe how strongly the given two random variables are related to each other.
It is a measure of correlation.	It is defined as the scaled form of covariance.
The value of covariance lies between -∞ and +∞.	The value of correlation lies between -1 and +1.
It indicates the direction of the linear relationship between the given two variables.	It measures the direction and strength of the linear relationship between the given two variables.

Data Analysis Process

- Everyone doing analysis has some missing data, especially survey researchers, market researchers, database analysts, researchers and social scientists.
- Missing data are questions without answers or variables without observations.

Data Analysis Process

- Even a small percent of missing data can cause serious problems with your analysis leading you to draw wrong conclusions. Real-world databases are highly susceptible to noise, missing, and inconsistent data due to they are typically huge in size often in gigabytes or more.
- We have to preprocess the data in order to help improve to quality of data and so as to improve the efficiency and ease of mining access. There are number of data preprocessing techniques.

Data Analysis Process

- Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube.
- Data transformations, such as normalization, may be applied. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance

Need of preprocessing data:

- The data you wish to analyze by data mining techniques are incomplete (lacking attribute values or certain attributes of interest), noisy (containing errors) and inconsistent.
- Incomplete data can occur in many reasons.
- Attribute values may not be available, not considering important at the time of entry.
- Missing data, particularly tuples with missing values for some attributes, may need to be inferred

Data cleaning:

- Real world data tend to be noisy, incomplete, and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.
- We concentrate mainly on filling of missing values by ignoring the data row completely, filling the missing values manually, use the global constant to fill the missing values, use the attribute mean for 1 column of data, same using to fill all columns of data, using most probable value to fill missing value (Regression algorithm).

Data cleaning with Regression Method

- In the regression method, a regression model is fitted for each variable with missing values.
- Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable. Since the data set has a several missing data patterns, the process is repeated sequentially for variables with missing values.

How to deal Missing Values?

- We have to fill those missing data cells with 6 possible ways.
- 1. Ignoring the data row completely
- 2. Filling missing values manually
- 3. Use a global constant to fill the missing values
- 4. Use the attribute mean to fill the missing value
- 5. Use the attribute mean for all samples belonging to the same class as the given tuple
- 6. Use the most probable value to fill the missing value (Predicting by Regression algorithm)

Regression Methodology

- A regression is a statistical analysis, assessing the association between two variables.
- It is used to find the relationship between two variables.
- RegressionFormula:
- RegressionEquation (y) = a + $\text{Intercept(a)} = \frac{\Sigma Y - b(\Sigma X)}{N}$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

  - slope -‘b’,
  - Intercept- ‘a’
  - r-coefficient of corrélation
  - x and y are the variables.

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

  - b = the slope of the regression line
  - a = the intercept point of the regression line and the y axis.
  - N = Number of values or elements
  - X = First Score
  - Y = Second Score
  - ΣXY = Sum of the product of first and Sec
  - ΣX = Sum of First Scores ΣY = Sum of Seco
  - ΣX^2 = Sum of square First Scores

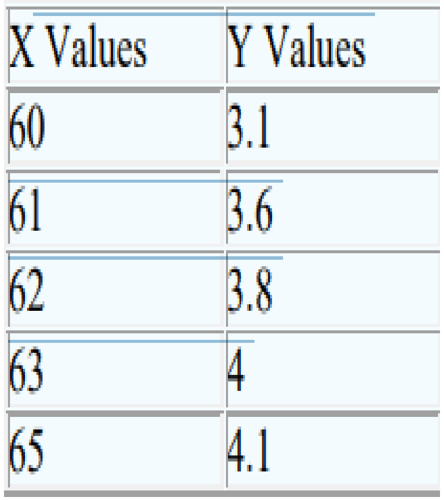
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Formula

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

How to find Linear Regression ?



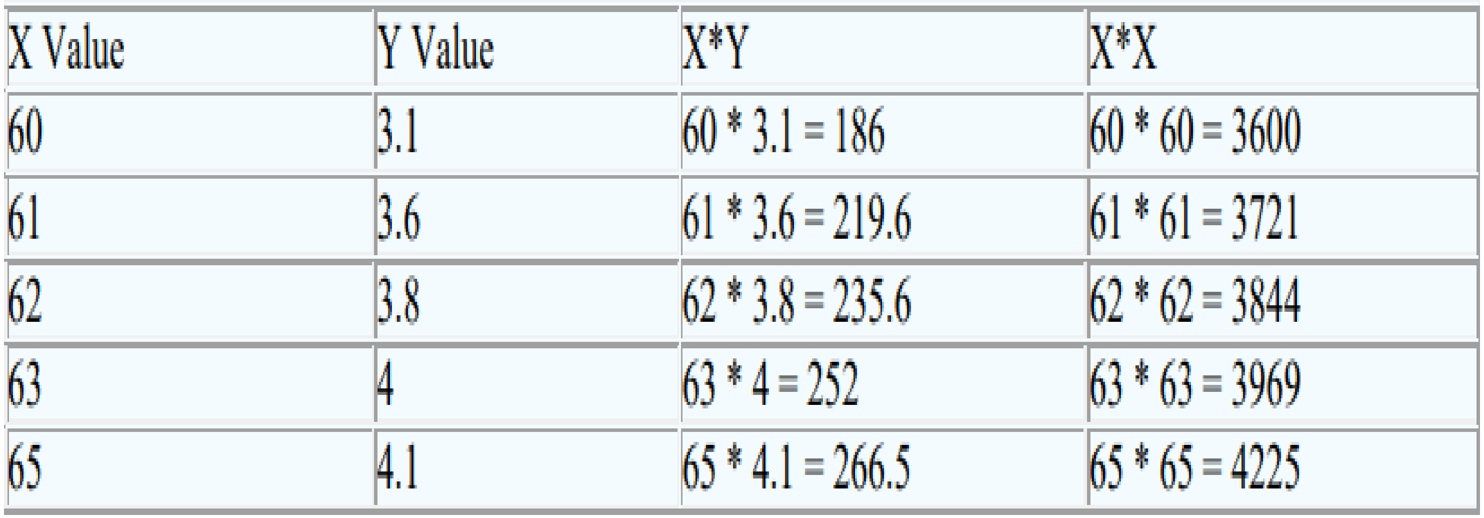
X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

First we will find slope,
intercept and
use it to form regression equation

Step by Step approach:

Step 1. Count the number of values. n = 5

Step 2: Find XY, X^2



X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

- Step 3: Find ΣX, ΣY, ΣXY, ΣX^2 .
  - ΣX = 311
  - ΣY = 18.6
  - ΣXY = 1159.7
  - ΣX2 = 19359
- Step 4: Substitute the above information in slope formula(

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

  - ((5)*(1159.7)- (311)*(18.6))/((5)*(19359)-(311)2)
  - = (5798.5 - 5784.6)/(96795 - 96721)
  - = 13.9/74
  - **b= 0.19**

- Step 5: Compute Intercept or a by fitting the value in the formula:
  - $\Sigma X = 311$
  - $\Sigma Y = 18.6$
  - $\Sigma XY = 1159.7$
  - $\Sigma X^2 = 19359$

$$\text{Intercept(a)} = \frac{\Sigma Y - b(\Sigma X)}{N}$$

  - Then
  - =(18.6 - 0.19(311))/5
  - = (18.6 - 59.09)/5
  - = -40.49/5
  - = -8.098

- Step 6: Then substitute these values in regression equation formula
- Regression Equation(y) = a+b(x)
- = -8.098 + 0.19x.
- Example of Prediction with Regression:
- Suppose if we want to know the approximate y value for the variable x = 64.
- Then we can substitute the value in the above equation.
- Regression Equation(y) = a + bx = -8.098 + 0.19(64).

= -8.098 + 12.16

= 4.06

Prediction of Value with Linear Regression

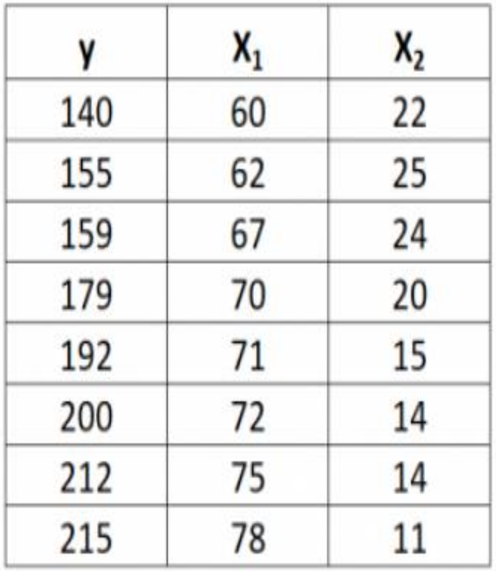
X Value	Y Value
60	3.1
61	3.6
62	3.8
63	4
65	4.1
64	4.06

Multiple linear regression

- **Multiple linear regression is a method we can use to quantify the relationship between two or more predictor variables and a response variable.**

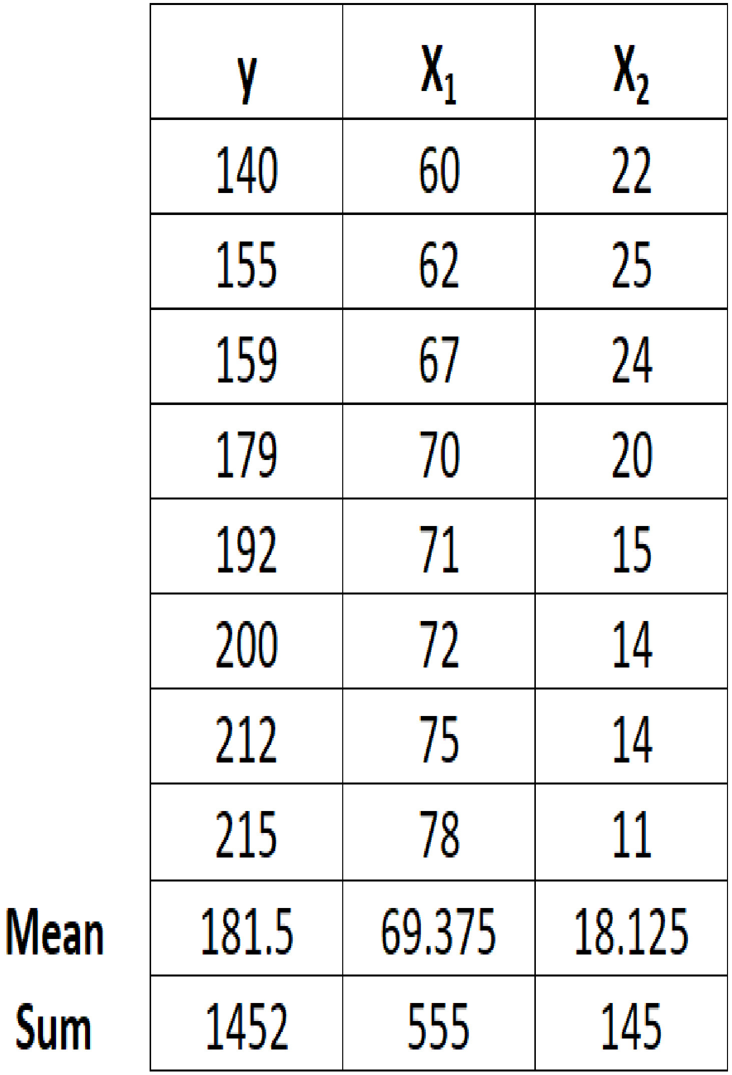
Multiple Linear Regression using Paper Pen

- Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2

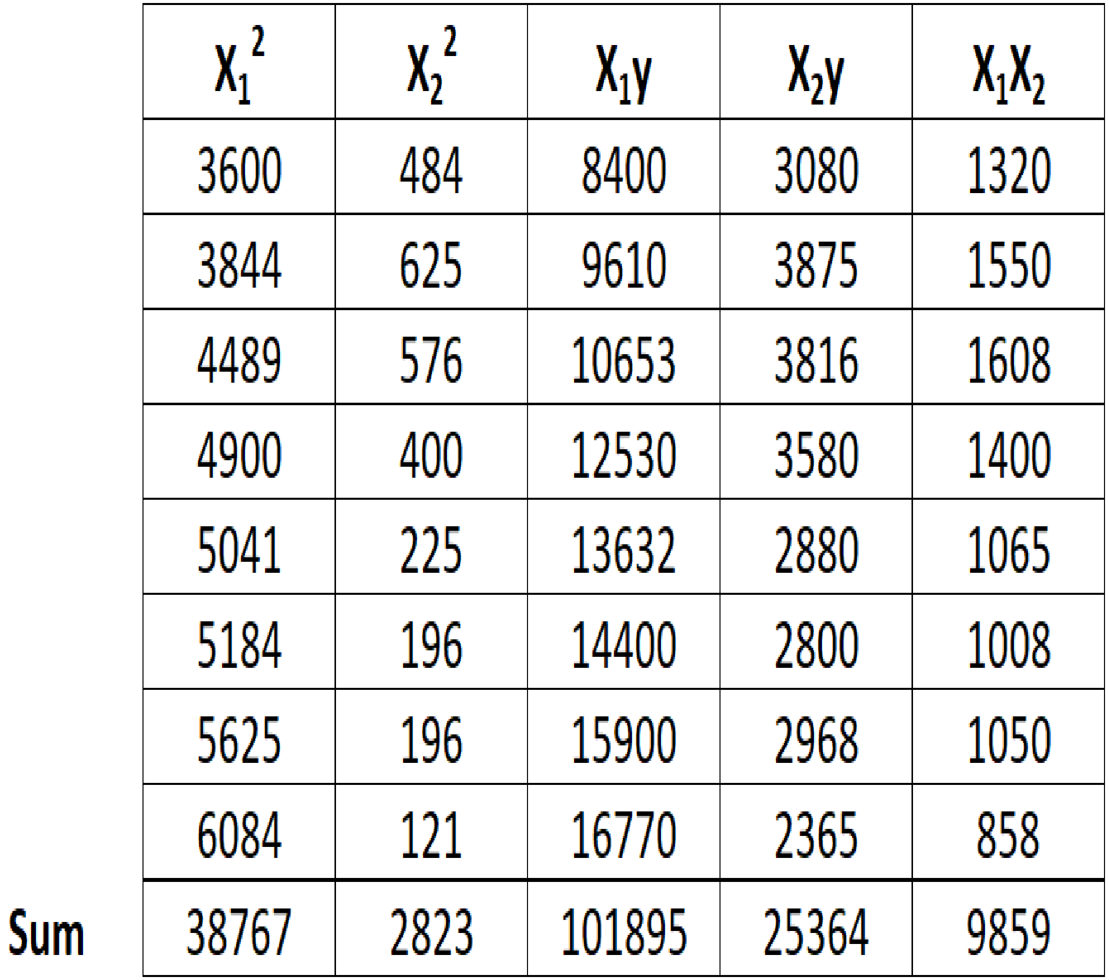


y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

- **Step 1: Calcúlate X_1^2, X_2^2, X_1y, X_2y and X_1X_2**



	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145



	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

- **Step 2: Calculate Regression Sums.**

- $\Sigma x_1{}^2 = \Sigma X_1{}^2 – (\Sigma X_1)^2 / n = 38,767 – (555)^2 / 8 =$ **263.875**
- $\Sigma x_2{}^2 = \Sigma X_2{}^2 – (\Sigma X_2)^2 / n = 2,823 – (145)^2 / 8 =$ **194.875**
- $\Sigma x_1y = \Sigma X_1y – (\Sigma X_1\Sigma y) / n = 101,895 – (555*1,452) / 8 =$ **1,162.5**
- $\Sigma x_2y = \Sigma X_2y – (\Sigma X_2\Sigma y) / n = 25,364 – (145*1,452) / 8 =$ **-953.5**
- $\Sigma x_1x_2 = \Sigma X_1X_2 – (\Sigma X_1\Sigma X_2) / n = 9,859 – (555*145) / 8 =$ **-200.375**

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

- **Step 3: Calculate b0, b1, and b2**
- The formula to calculate b_1 is:
- $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$
- Thus, $\mathbf{b_1}$ = [(194.875)(1162.5) – (-200.375)(-953.5)] / [(263.875) (194.875) – $(-200.375)^2$]
- = **3.148**
- The formula to calculate b_2 is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$
- Thus, $\mathbf{b_2}$ = [(263.875)(-953.5) – (-200.375)(1152.5)] / [(263.875) (194.875) – $(-200.375)^2$]
- = **-1.656**
- The formula to calculate b_0 is: $y - b_1X_1 - b_2X_2$
- Thus, $\mathbf{b_0}$ = 181.5 – 3.148(69.375) – (-1.656)(18.125)
- = **-6.867**

- **Step 4: Place b_0, b_1, and b_2 in the estimated linear regression equation.**
- The estimated linear regression equation is:
- $\hat{y} = b_0 \quad + b_1 * x_1 \quad + b_2 * x_2$
- $\boldsymbol{\hat{y} = -6.867 + 3.148x_1 - 1.656 \; x_2}$
-
- **Example of prediction:**
  Suppose Value x1= 76 and x2=13 then value of y ?
- **= -6.867 + 3.148 * 76 – 1.656 *13**
- **=239.248-21.528-6.867**
- **239.248-28.395=210.853**
- **Additionally,**
- **-6.867 + 3.148 * 70 – 1.656 *20**
- **=180.373**

Thanks!!!!