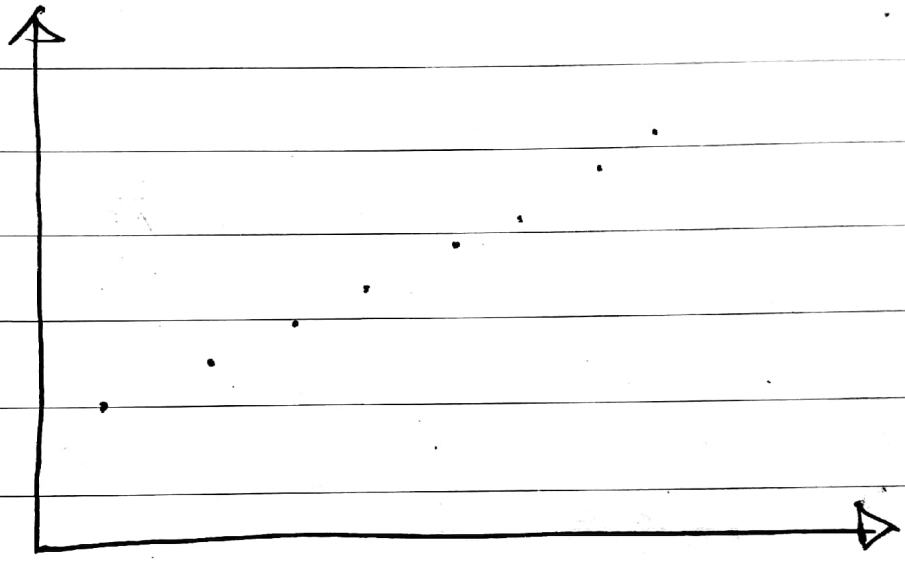
CSE - 313

Notes

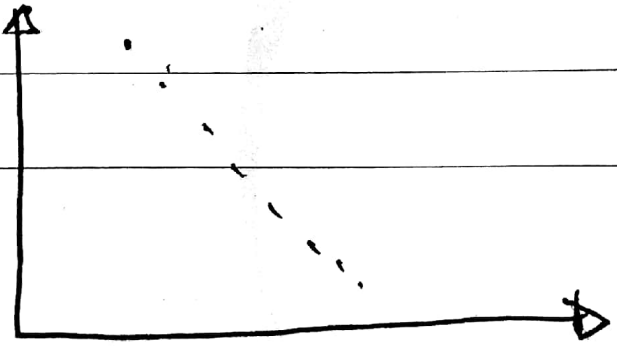
Correlation: is usually defined as a measure of the linear relationship between two quantitative variables (e.g: height and weight)

→ is a statistical measure that indicates the extent to which two or more variables fluctuate together.

* ~~Posi~~ A positive correlation indicates the extent to which those variables increase or decrease in parallel.



* A negative correlation indicates the extent to which one variable increases as the other decreases.



→ A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.

Pearson's Product moment Coefficient:

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n \sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{n} (a_i b_i) - n \bar{A} \bar{B}}{n \sigma_A \sigma_B}$$

Here,

n = no. of tuples

$\bar{A}$ and $\bar{B}$ are the respective means

σ_A and σ_B are the respective S.D of A and B

→ If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.

→ $r_{A,B} = 0$; independent ;

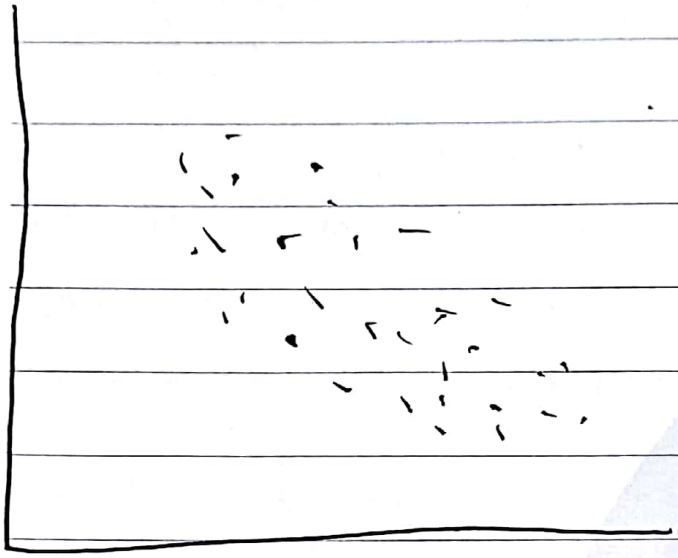
→ There is no correlation

$r_{A,B} < 0$: negatively correlated

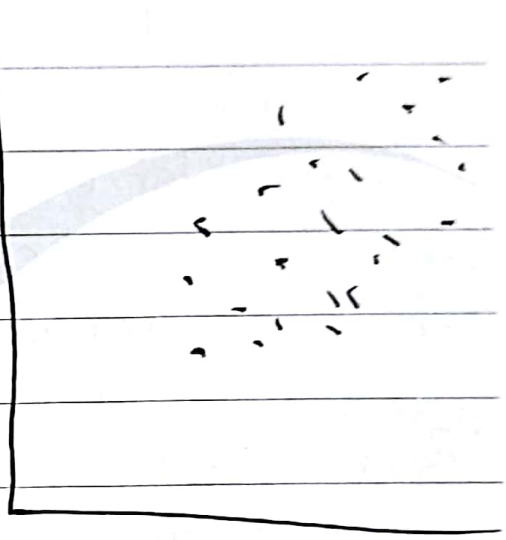
Notes

~~One attribute~~

- A correlation of 1 indicates a perfect positive correlation.
- A correlation of -1 . . . . . perfect negative correlation



Weak negatively correlated data



Weak positively correlated da

age	23	23	27	27	39	41
% fat	9.5	~~7.8~~ 26.5	~~17.~~ 8	~~31.4~~ 17.8	31.4	25.9
47	49	50	52	54	54	56
~~27.4~~	27.2	31.2	34.6	42.5	28.8	33.4

Regression:

There is single response variable Y, also called the dependent variable, which depends on the value of a set of input, also called independent variables $x_1, x_2, \ldots x_n$.

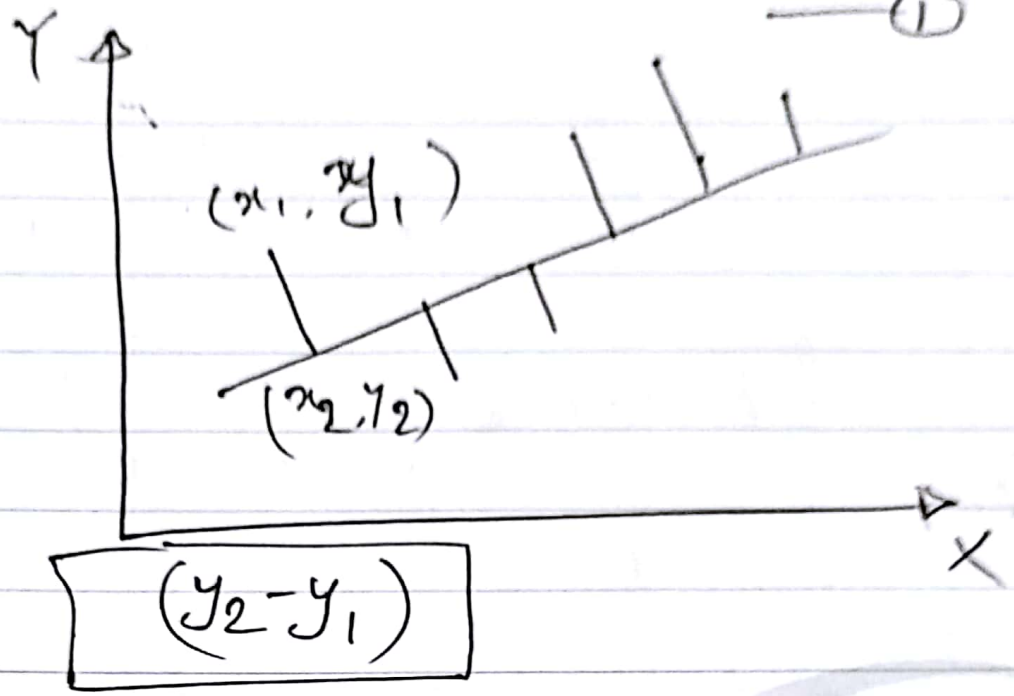
→ A technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with and depends on, a change in one or more independent variables.

* Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest.

The simplest type of relationship between the dependent variable Y and the input variables $x_1, \ldots, x_n$ is a linear relationship for some constants $\beta_0, \beta_1, \ldots \beta_n$.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad \text{---(I)}$$

Notes



ϵ = random error

For error,

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + e \quad \text{---(II)}$$

↳ is called a linear regression equation. $\beta_0, \beta_1, \ldots \beta_n$ are called regression coefficients

Q: Difference between Correlation and Regression. (Next class)

↳ Self study

Notes

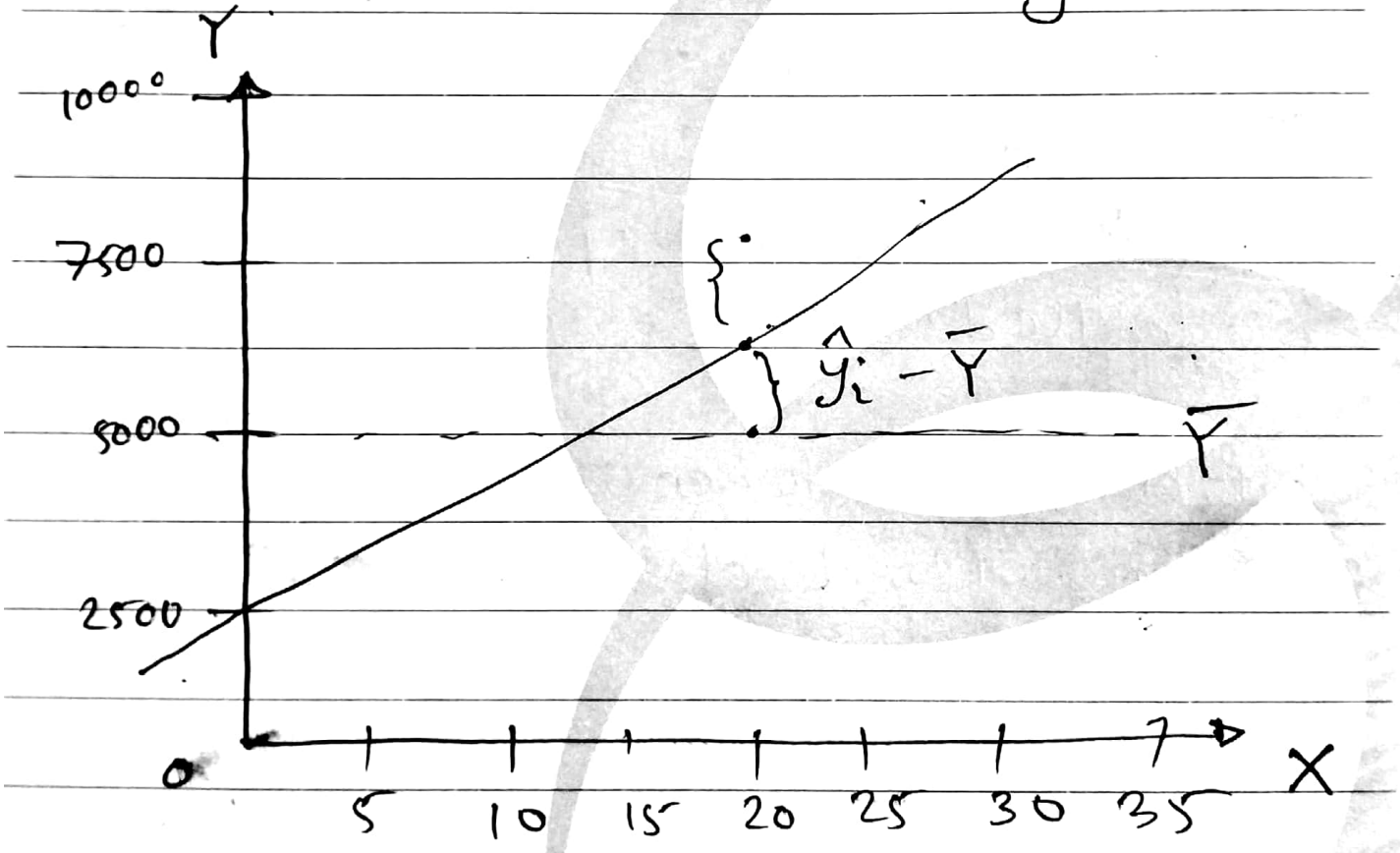
Simple linear Regression: A regression equation containing a single independent variable (that is, one in which $n=1$) is called a simple linear regression equation.

$Y = \beta_0 + \beta_1 x_i + \epsilon$ → Population regression function

β_0 → unknown intercept

β_1 → Slope parameter

multiple linear Regression: A regression equation containing many independent variables is called a multiple linear regression equation



Notes

Explained $= \hat{Y}_i - \bar{Y}$

Unexplained $= Y_i - \hat{Y}_i$

SSR = Sum of squares due to regression

$= \Sigma (\hat{Y}_i - \bar{Y})^2$

SSE = Sum of squares due to error/residual

$= \Sigma (Y_i - \hat{Y}_i)^2$

SST = Sum of squared total

$= SSR + SSE$

$= \Sigma (Y_i - \bar{Y})^2$

$R^2 = \frac{SSR}{SST}$

→ Proportion of total variation that is explained

Linear Regression function:

$y = a + bx$

Slope of Regression line:

$b = r \cdot \frac{S_y}{S_x}$

r = Pearson's Correlation Coefficient

$= \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}}$

Notes

Y - intercept of regression line :

$$a = \bar{Y} - b\bar{x}$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
17	94					
13	73					
12	59					
15	80					
16	93					
14	85					
16	66					
16	79					
18	77					
19	91					

$\bar{x} =$ $\bar{y} =$

1) Find the regression eqn and draw the regression line

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} \qquad S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

The co-efficient of determination & the sample correlation coefficient :

To measure the amount of variation in the set of response variables

$Y_1, Y_2, \ldots, Y_n$ of the amount of variation in a set of values $Y_1, \ldots, Y_n$ is given by,

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

Y_i = observed

$\hat{Y}$ = predicted

$$R^2 = \frac{SSR}{SST}$$

R^2 represents the proportion of the variation in the response variables that is explained by the different types of input values. It's called the coefficient of determination.

$0 \leq R^2 \leq 1$

If $R^2 \to 1$, it indicates that most of the variation of the response data is explained by the different input values.

$R^2 \to 0$, it indicates that little of the variation is explained by the different input values.

Now the sample correlation co-efficient

Notes

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

As r is a unitless coefficient, it provides a measure of the degree to which high value of x are paired with high values of y and low values of x with low values of y.

$$-1 \leq r \leq 1$$

$r \longrightarrow 1$ strongly positively correlated

$r \longrightarrow -1$ large x values are strongly associated with small y values and small x values with large y values.

Thus $(r)^2$ = similar to R^2

$$\therefore |r| = \sqrt{R^2}$$

Q: The grades of a class of 9 students on a midterm report (x) and on the final exam (y) are as follows:

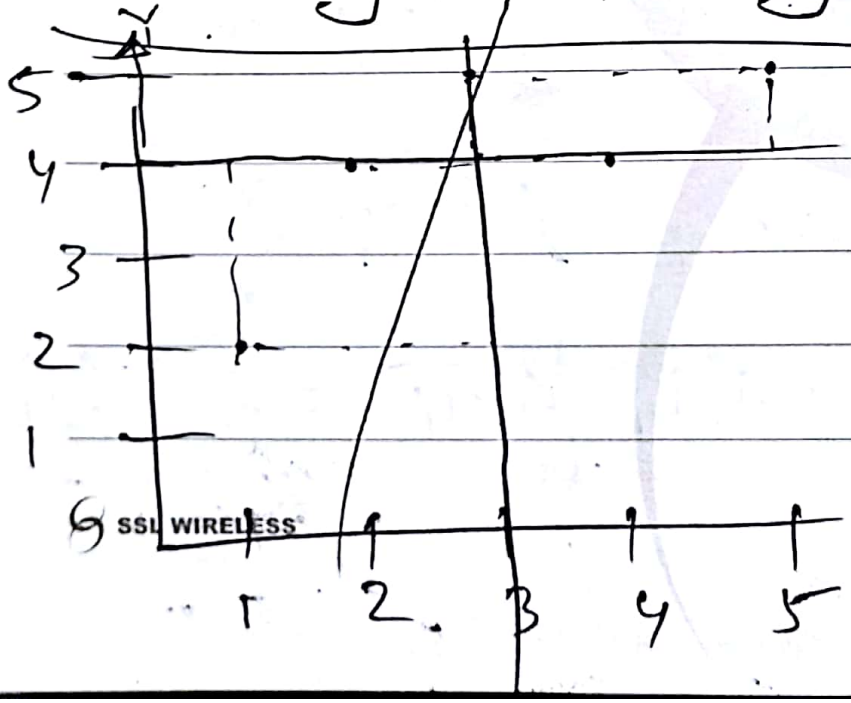
Notes

x	y
85	75
82	95
71	78
72	84
81	47
94	8.5
96	99
99	99
67	68

a) Estimate the line

b) Estimate the final exam grade of a student who received a grade of 85 on the midterm report.

Linear regression using least square method



x	y
1	2
2	4
3	5
4	4
5	5
mean	3

Least Square method:

Notes

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\frac{\partial(SSE)}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial(SSE)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$$

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \checkmark \quad (i)$$

Again,

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (ii)$$

Notes

$n b_0 + b_1 \sum_{i=1}^{n} x_i$

$$n b_0 = \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i$$

$$\Rightarrow b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}$$

$$= \bar{y} - b_1 \bar{x}$$

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i$$

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \ldots}{\ldots}$$

(1) $\times n$ $\quad$ (1) $\times \sum x_i$

$$n b_0 \sum_{i=1}^{n} x_i + n b_1 \sum_{i=1}^{n} x_i^2$$

$$- n b_0 \sum_{i=1}^{n} x_i + b_1 \left(\sum_{i=1}^{n} x_i \right)^2 = n \sum_{i=1}^{n} x_i y_i$$

$$- \sum_{i=1}^{n} x \sum_{i=1}^{n} y_i$$

Notes

$$\Rightarrow b_1 \left\{ n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right\}$$

$$= n \sum x_i y_i - \sum x_i \sum y_i$$

$$\Rightarrow b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2}$$

(1) $\Rightarrow$

$$b_0 = \frac{\sum y_i - b_1 \sum x_i}{n}$$

$$= \bar{y} - b_1 \bar{x}$$