L- 8: Predictive Analytics

Agenda

- Covariance
- Correlation
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

$$\mathrm{Cov}(X,Y) =$$

$$= \left[E\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]$$

$$= \sum\sum \left(x-\mu_x\right)\left(y-\mu_y\right) p(x,y)$$

if discrete

$$= \iint \left(x-\mu_x\right)\left(y-\mu_y\right) f(x,y)\, dx\, dy$$

if continuous

consider the following

⇒ whether spending on advertising of a company is related to overall sales of the company.

→ If it is related, how it is related

⇒ Forecasting the sales, given the budget for advertising

And also

⇒ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

We need to validate this?

How ⟶ ?

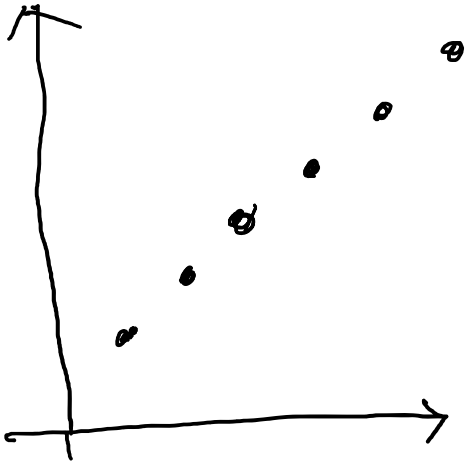
Correlation

→ Sales of a company and Expenditure on advertisement

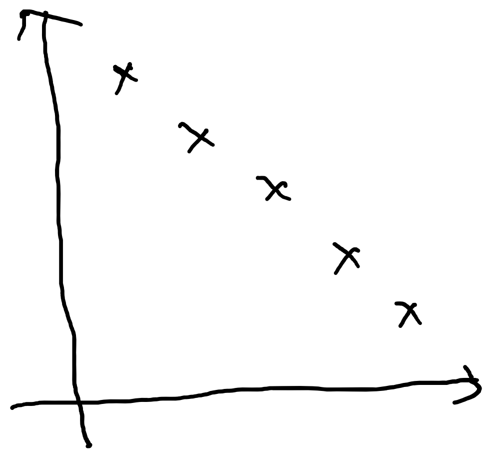
→ Price and Demand of a product

→ Inflation and Gold Price

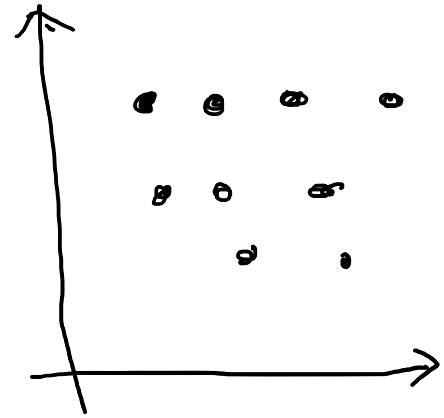
→ IQ and performance in Entrance.



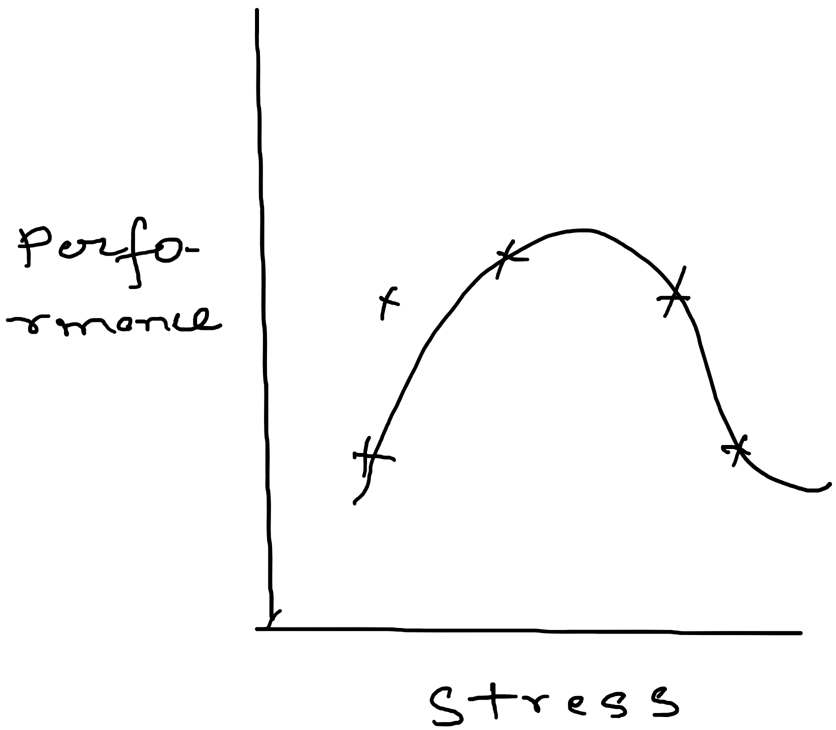
positive correlation



Negative correlation



No correlation


Perfo-
rmance
Stress


results
efforts

Coefficient of correlation:

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \, \sigma_y} = \frac{\Sigma X Y}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$

where

$$X = x - \bar{x}$$

$$Y = y - \bar{y}$$

$$X^2 = (x - \bar{x})^2$$

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

Coefficient of Correlation

$r = 1 \Rightarrow$ Perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ no relation

$0 < r < 1 \Rightarrow$ partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

Example - 1

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

r

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	-4	16	10	-4	16	16
2	-3	9	11	-3	9	9
3	-2	4	12	-2	4	4
4	-1	1	14	0	0	0
5	0	0	13	-1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	Σ 59

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

Coefficient of Determination

r is coeff. of Correlation

r^2 is coeff of determination

↓

Indicates the extent to which variation in one variable is explained by the variation in the other.

$r = 0.9 \Rightarrow r^2 = 0.81$

i.e 81% of the variation in y due to variation in 'x'

remaining 19% is due to some other factors.

Regression

Regression :-

x	1	2	3	4	5
y	1	4	9	16	25

when $x = 7$: $y = ?$

x	1	2	3	4	5
y	1	6	2	5	4

when $x = 7$, $y = ?$

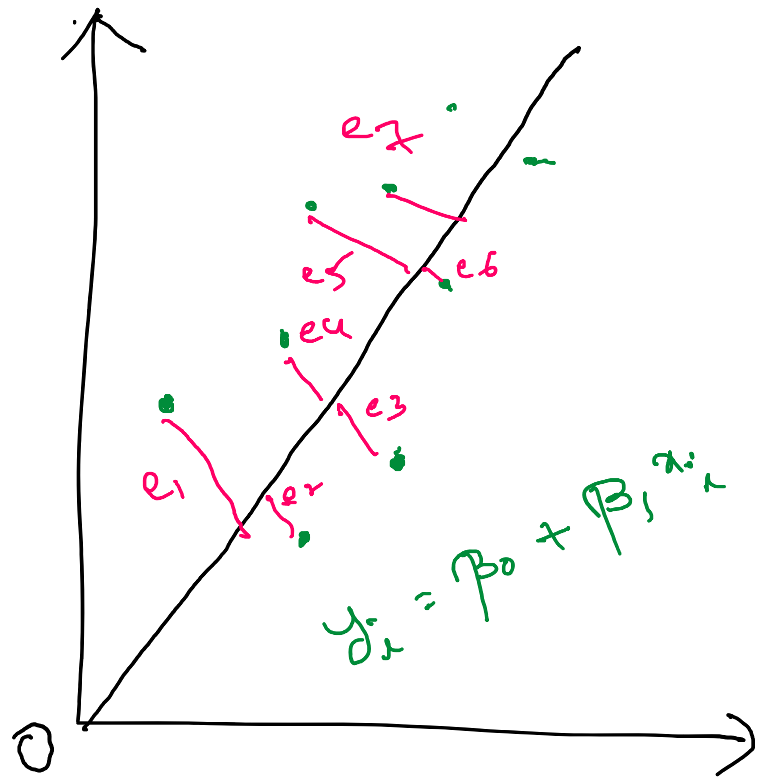
Method of Least squares

y : Dependent Variable

x : Independent Variable

predictor variable

response Variable

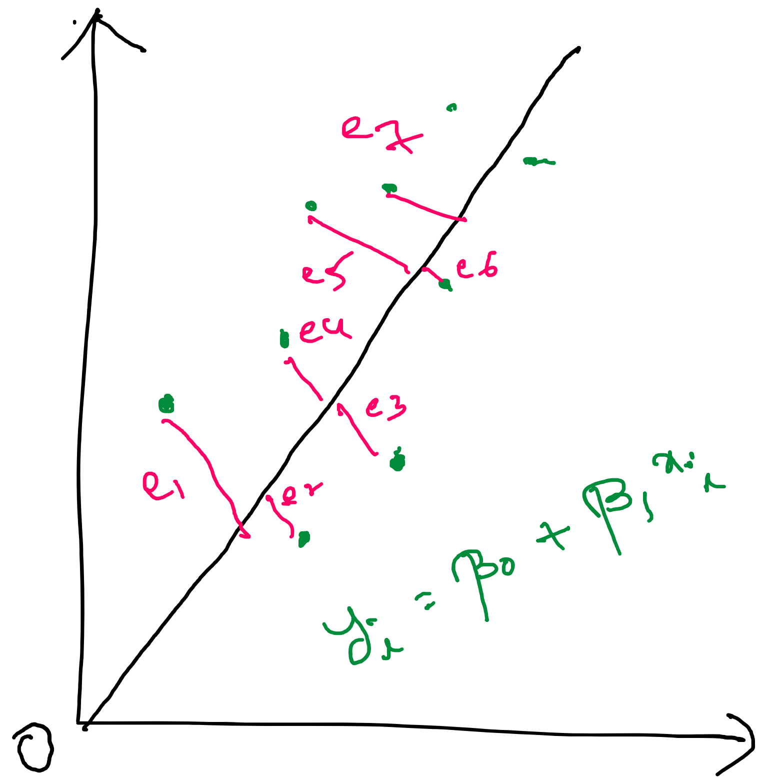


Method of Least squares

$S(\beta_0, \beta_1)$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

β_0 and β_1 are chosen to minimize the sum of squared errors.



Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)(-1)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)(2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression

$$y = \beta_0 + \beta_1 x$$

$$\left.\begin{aligned} \sum y &= \beta_0 n + \beta_1 \sum x \\ \sum xy &= \beta_0 \sum x + \beta_1 \sum x^2 \end{aligned}\right\}$$

Normal equations.

Matrix Approach:

Let $y = \beta_0 + \beta_1 x_1 + \cdots\cdots + \beta_{p-1} x_{p-1}$

observations $y_i = 1, 2, \ldots n \rightarrow$ by a vector Y

unknowns $\beta_0, \beta_1 \cdots\cdots \beta_{p-1} \rightarrow$ " " β

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{bmatrix}$$

$$\underset{n\times 1}{\hat{Y}} = \underset{n\times p}{X} \; \beta_{p\times 1}$$

Find β to minimize

$$S(\beta) = \sum_{i=1}^{m} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots \right)^2$$

$$= \| Y - X\beta \|^2 = \| Y - \hat{Y} \|^2$$

Diff S wrt to each β we get Linear eqns

$$X^T X \hat{\beta} = X^T Y \longrightarrow \text{normal eqns}$$

If $X^T X$ is non-singular, the soln is

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y$$

Computationally, it is sometimes unwise even to form the normal equations because the multiplications involved in forming $X^T X$ can introduce undesirable round-off error.

Linear regression
(Multiple regression)

Example:-

x_0	size	No of rooms	No of floors	Age of home	price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000
	x_1	x_2	x_3	x_4	y

Linear regression
(Multiple regression)

Example:-

$$X = \begin{bmatrix} 1 & 2000 & 5 & 2 & 45 \\ 1 & 1400 & 3 & 1 & 40 \\ 1 & 1600 & 3 & 2 & 30 \\ 1 & 800 & 2 & 1 & 35 \end{bmatrix} \qquad y = \begin{bmatrix} 4000 \\ 2000 \\ 3000 \\ 2000 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T y$$

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma\, 3.5$	$\Sigma\, 10.5$	$\Sigma\, 21$

$$y = \beta_0 + \beta_1 x$$

$$\Sigma y = n\beta_0 + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$$

$$3.5 = 4\beta_0 + \beta_1 (7)$$

$$10.5 = 7\beta_0 + \beta_1 (21)$$

on solving these

$$\beta_0 = 0$$

$$\beta_1 = 0.5$$

ie $y = 0 + (0.5)x$

$$\boxed{\text{when } x = 5, \quad y = (0.5)5 = 0.25}$$

Thanks