Lecture 5

Regression

→ **Regression Types**

Based on types of functions used.

1) Simple Linear Regression
2) Multiple Linear Regression
3) Polynomial Regression
4) Logistic Regression

→ Predict/understanding/finding a relationship between one or more independent and one or more dependent variable.

Example:

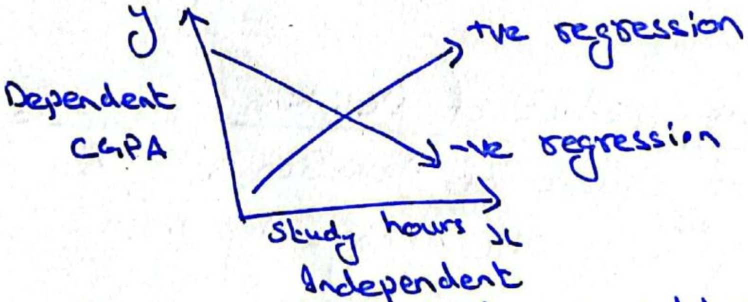
1) Predict height of a person given his age.
2) Predict car price given model/year/mileage/engine capacity.
3) Salary based on years of experience/education.

Given

x = Independent variable

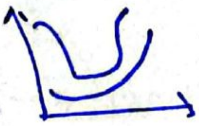
y = dependent variable (variable being predicted)

Predict exam score based on study hours.



→ When multiple independent variables, that is known as Multiple Linear Regression.

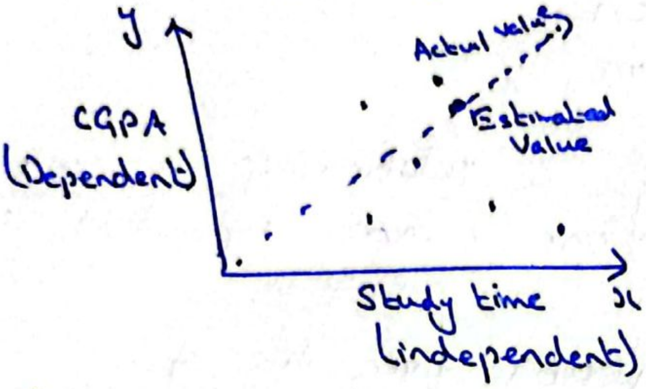
→ When line is not straight, it is polynomial regression.



Linear Regression:

$$\underset{\substack{\text{Predicting}\\ \text{Dependent}}}{y} = \underset{\text{Slope}}{m}\overset{\text{independent}}{x} + \underset{\text{intercept}}{b}$$

m = How much y changes for a unit change in x.



Regression line based on "Least Squared" method.

Example:

Pizza Diameter (x)	Price (y)	Mean (x) $\frac{x_1+x_2+x_3}{3}$	Mean (y) $\frac{y_1+y_2+y_3}{3}$	Deviation (y) y - Mean	Product of Deviation	Sum of Product of Deviation	Square of Deviation for x ~~(Dev of x)~~^{2}	Deviation of (x)
8	10	10	13	-3	6	12	4	-2
10	13			0	0		0	0
12	16			3	6		4	2

$$\text{Calculate } m = \frac{\text{Sum of product of Deviation}}{\text{Sum of square of Deviation } x} = \frac{12}{8} = 1.5$$

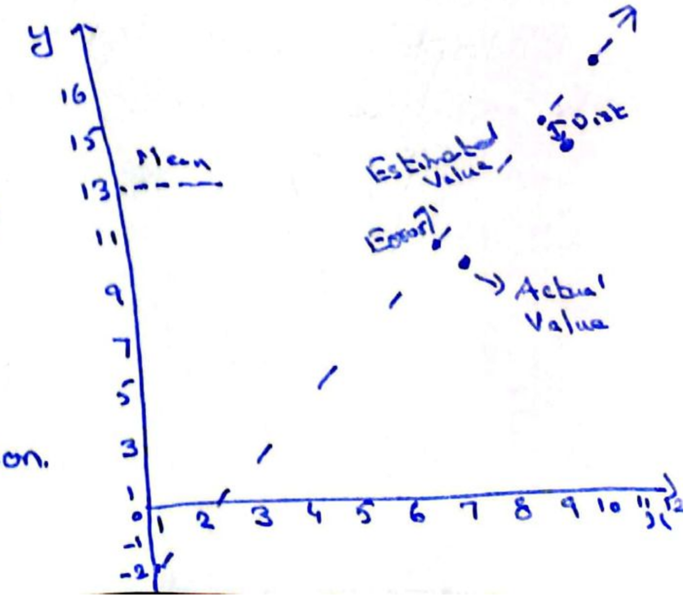
→ if you change x by '1', y will change by 1.5.

Calculate b = Mean of y − (m × Mean of x)

$$= 13 - (1.5 \times 10)$$
$$= 13 - 15$$
$$= -2$$

→ So suppose if someone ask, what will be the price of 20' pizza.

$$y = mx + b$$
$$= (1.5 \times 20) + (-2)$$
$$= 30 - 2 = \boxed{28} \text{ prediction.}$$



or.

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

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

→ Best case scenario, when all your values are on straight line.
When you have a lot of data, the points will be scattered and not on line.

→ R^2 is the percent of 'y' variation explained by 'x'.

→ It tells us, how accurately the regression line predicts or estimates the actual value.

→ Distance (actual - mean)
→ Distance (estimated - mean)

$\bar{y}$ = Mean of y
$\hat{y}$ = Estimated value.

$\bar{y} = 13$
$\hat{y} = -2 + 1.5x$

$y - \bar{y}$	$(y-\bar{y})^2$	Est value $\hat{y}$	Distance betw $\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
-3	9	10	-3	9
0	0	13	0	0
3	9	16	3	9
	18			18

$$R^2 = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2} = \frac{18}{18} = 1 \text{ (perfect).}$$

→ Multiple Linear Regression :-

- → In Linear regression, 1 dependent & 1 independent Variable.
- → In Multiple LR, 1 dependent & multiple independent variables.
- → MLR of two variables x_1 & x_2 is given as;

$$y = f(x_1, x_2)$$

$$y = a_0 + a_1x_1 + a_2x_2$$

In general, for 'n' independent variables

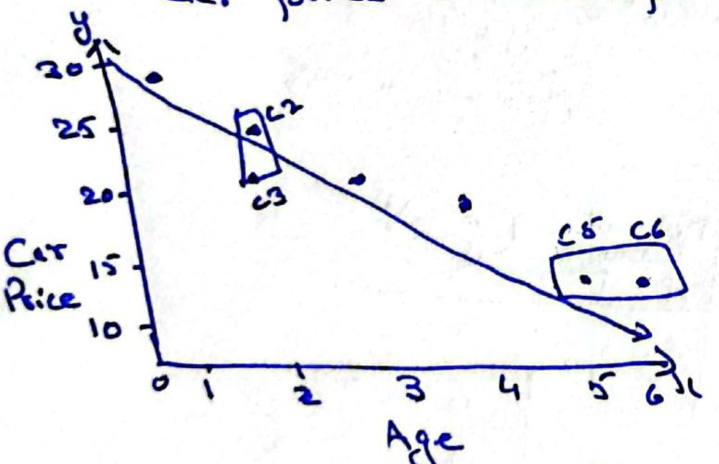
$$y = f(x_1, x_2, \ldots\ldots, x_n)$$

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + \varepsilon$$

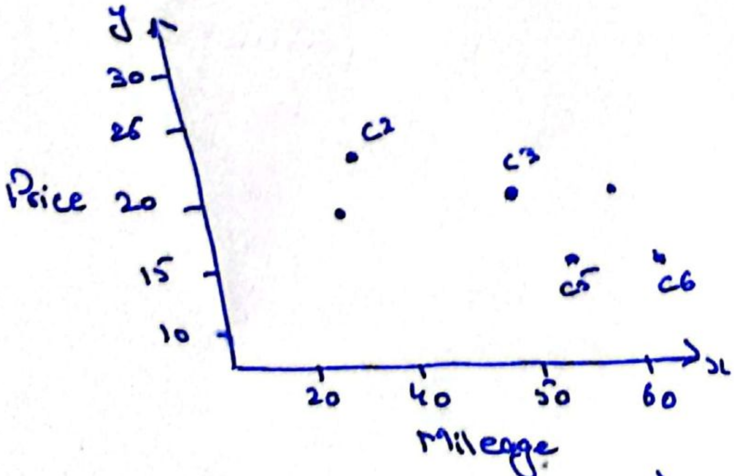
Example:

Car price = intercept + Age + Milgge

Car	Price (k)	Years Age	Milegge (k)
1	29	1	18
2	25	2	25
3	21	2	50
4	18	3	68
5	15	4	75
6	15	5	65



$$\underset{y}{\text{Price}} = \underset{b}{30.57} + \underset{m}{(-3.55)}\,\underset{x}{\text{Age}} \rightarrow ①$$



$$\text{Price} = 32.04 + (-0.23) \cdot \text{Milegge} \longrightarrow ②$$

→ Now plot both these Linear regressions on the same figure to have a 2D plot.

Combining ① & ②

$$\text{Price} = \underbrace{34.46}_{\text{Brand new Car Price}} + (-1.54)\,\text{Age} + (-0.15)\,\text{Mileage}$$

- Age results in 10 times more in price reduction as compared to mileage.

10k miles $\cong$ 1.54 years.

- \$1.54k reduction with each year.
- \$0.15k reduction with each thousand miles.

eg. Car Age = 2
Mileage = 50k miles

$$\text{Price} = 34.46 - 1.54\,\text{Age} - 0.15\,\text{Mileage}$$
$$= 34.46 - 1.54(2) - 0.15(50)$$
$$= \$\,21.88\text{k}.$$

Numerical Example:

→ Matrices for x & y

$$x = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{bmatrix} \quad \& \quad y = \begin{bmatrix} 1 \\ 6 \\ 8 \\ 12 \end{bmatrix}$$

Product 1 x_1	Product 2 x_2	Weekly Sales y
1	4	1
2	5	6
3	8	8
4	2	12

Coefficient of MLR is

$$\hat{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

→ Calculate same as linear regression.

$$\hat{a} = ((X^TX)^{-1} X^T)Y$$

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 4 & 5 & 8 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 8 \\ 1 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 10 & 19 \\ 10 & 30 & 46 \\ 19 & 46 & 109 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 3.15 & -0.59 & -0.3 \\ -0.59 & -0.2 & 0.016 \\ -0.3 & 0.016 & 0.054 \end{bmatrix}_{3\times 3}$$

$$(X^T.X)^{-1}.X^T = \text{"} \begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 2 & 3 & 4\\ 4 & 5 & 8 & 2\end{bmatrix}$$

$$((X^T.X)^{-1}X^T).Y = \begin{bmatrix}0.05 & 0.47 & -1.02 & 0.19\\ -0.32 & -0.098 & 0.155 & 0.26\\ -0.065 & 0.005 & 0.185 & -0.125\end{bmatrix} . \begin{bmatrix}1\\ 6\\ 8\\ 12\end{bmatrix}$$

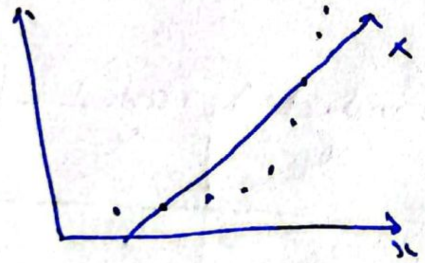
$$= \begin{array}{cl} \left[\begin{matrix}-1.69\\ 3.48\\ -0.05\end{matrix}\right] & \begin{matrix}a_0\\ a_1\\ a_2\end{matrix}\end{array}$$

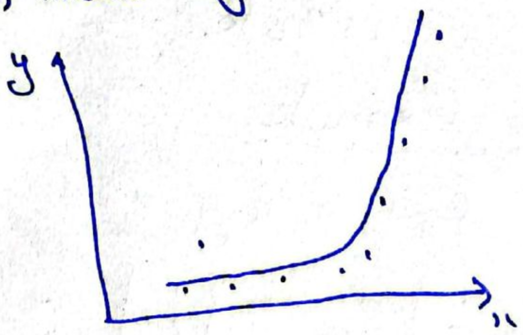
Hence

$$y = -1.69 + 3.48x_1 + (-0.05)x_2$$

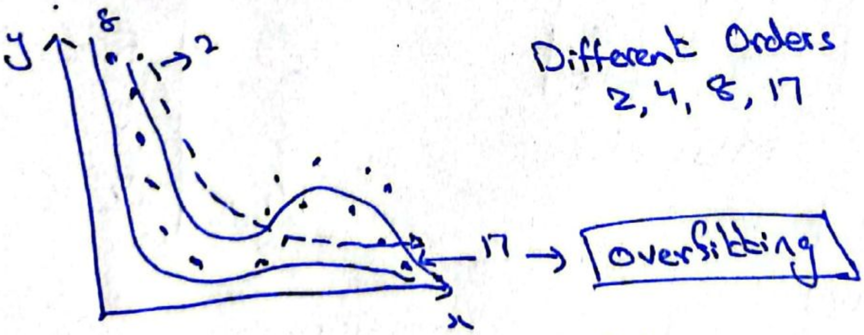
→ **Polynomial Regression:**

→ If the relationship between independent and dependent variables is not linear, linear regression will result in large errors.





Polynomial Regression



Different Orders 2, 4, 8, 17

→ We can use non-linear relationship among variables by using n^{th} degree of polynomial.

→ for example

$y = a_0 + a_1x + a_2x^2$ ⟶ second degree

$y = a_0 + a_1x + a_2x^2 + a_3x^3$ ⟶ Third degree

Numerical Example:

→ For 2nd degree $y = a_0 + a_1 x + a_2 x^2$ where coefficients a_0, a_1, a_2 are calculated using

x	y
1	1
2	4
3	9
4	15

$$a = X^{-1} B$$

where

$$X = \begin{bmatrix} n & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{bmatrix}^{-1} \quad B = \begin{bmatrix} \Sigma y_i \\ \Sigma (x_i, y_i) \\ \Sigma (x_i^2, y_i) \end{bmatrix}$$

x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y$	x_i^3	x_i^4
1	1	1	1	1	1	1
2	4	8	4	16	8	16
3	9	27	9	81	27	81
4	15	60	16	240	64	256

$\Sigma x_i = 10$; $\Sigma y_i = 29$; $\Sigma x_i y_i = 96$; $\Sigma x_i^2 = 30$; $\Sigma x_i^2 y_i = 338$

$\Sigma x_i^3 = 100$; $\Sigma x_i^4 = 354$

$$a = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}^{-1} \begin{bmatrix} 29 \\ 96 \\ 338 \end{bmatrix}$$

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -0.75 \\ 0.95 \\ 0.75 \end{bmatrix}.$$

$$y = -0.75 + 0.95x + 0.75x^2$$