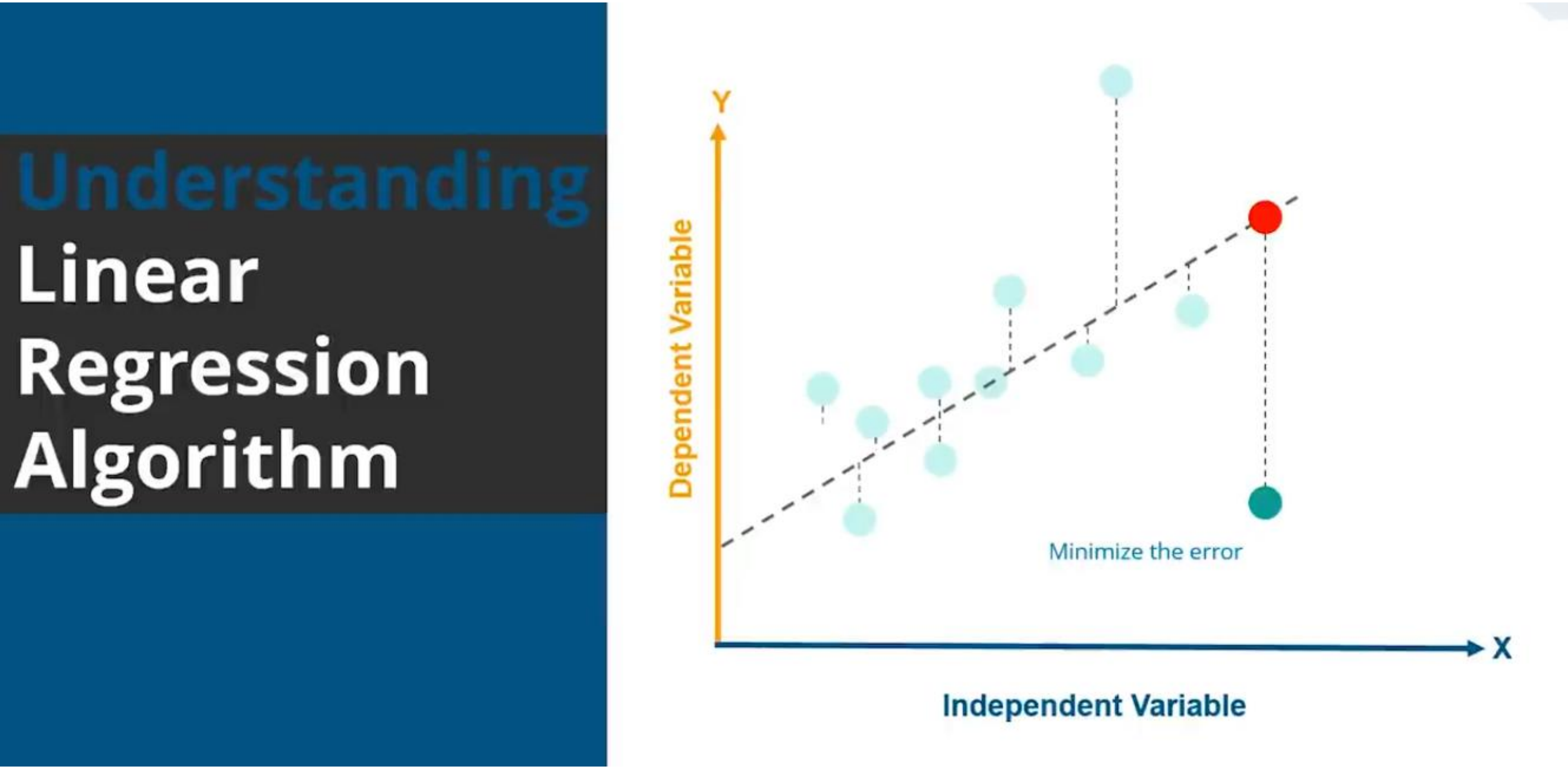
What is Regression?

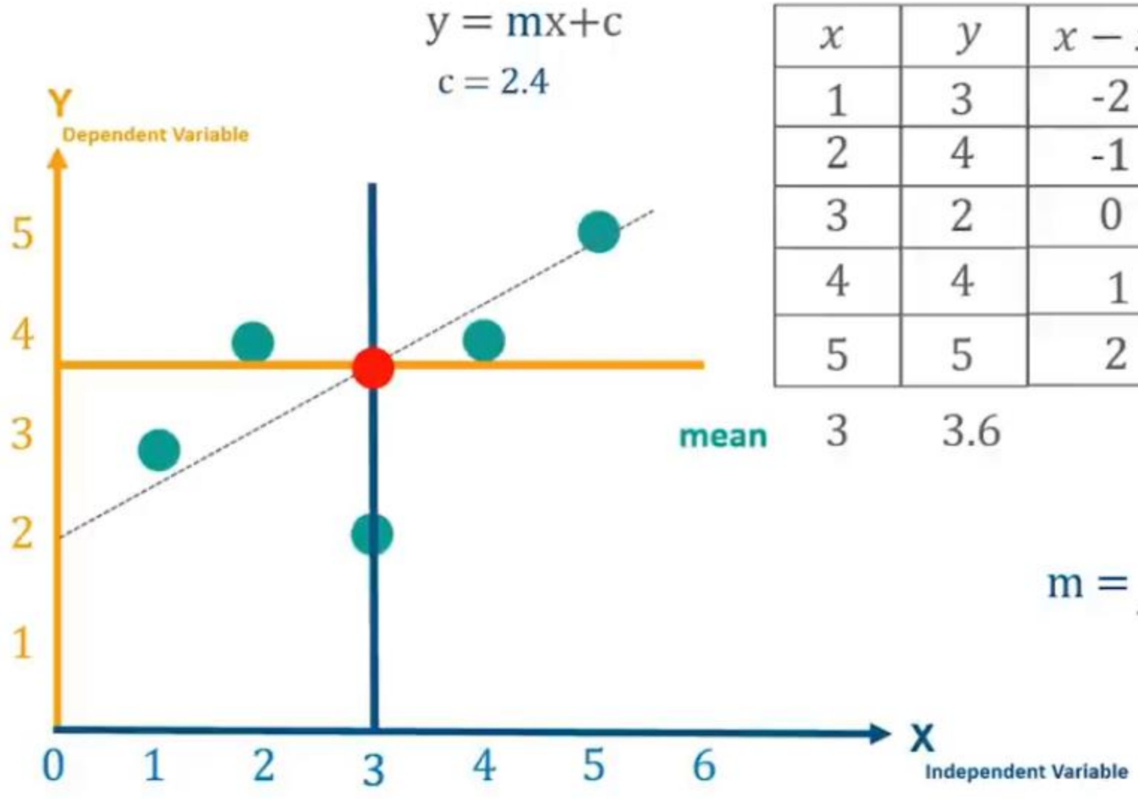
"Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable"

Where is Linear Regression used?

- Evaluating Trends and Sales Estimates
- Analyzing the Impact of Price Changes
- Assessment of risk in financial services and insurance domain


Understanding
Linear
Regression
Algorithm
Y
Dependent Variable
Minimize the error
X
Independent Variable

Understanding Linear Regression Algorithm



x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
3	3.6			$\Sigma = 10$	$\Sigma = 4$

mean: 3, 3.6

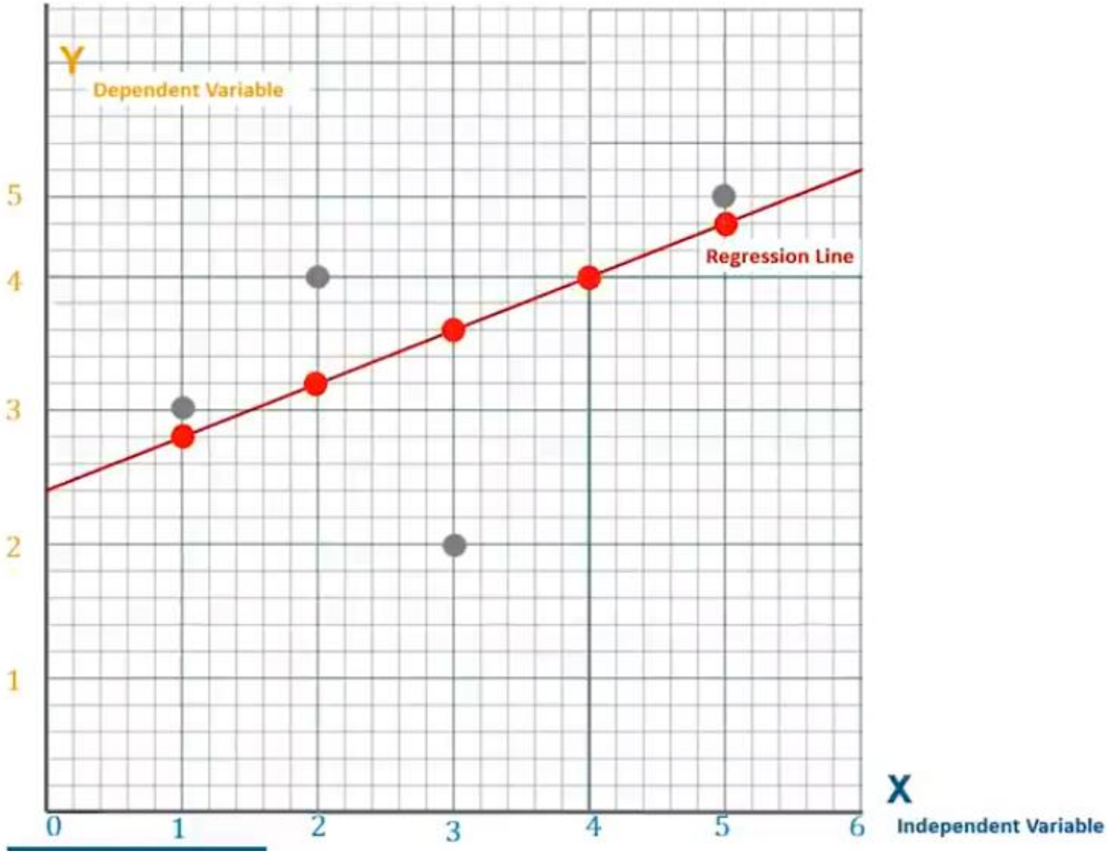
$$m = \sum \frac{(x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4 \quad c = \bar{y} - m\bar{x}$

$y = 0.4x + 2.4$

Mean Square Error

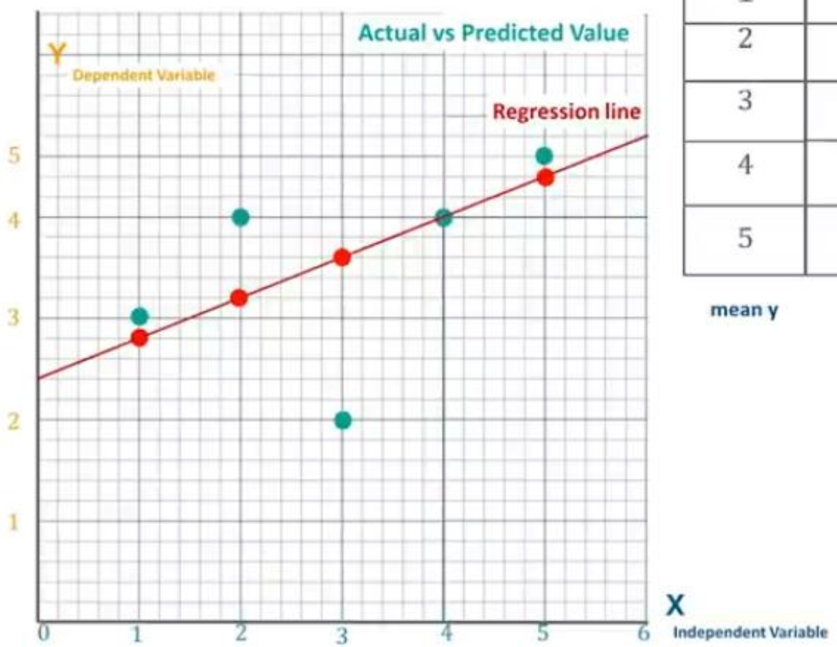


$m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$y = 0.4 \times 1 + 2.4 = 2.8$
$y = 0.4 \times 2 + 2.4 = 3.2$
$y = 0.4 \times 3 + 2.4 = 3.6$
$y = 0.4 \times 4 + 2.4 = 4.0$
$y = 0.4 \times 5 + 2.4 = 4.4$

Calculation of R^2



x	y	$y-\bar{y}$	$(y-\bar{y})^2$	y_p	$(y_p - \bar{y})$	$(y_p-\bar{y})^2$
1	3	− 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	−1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		Σ 5.2			Σ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\Sigma\left((y_p-\bar{y}\right)^2}{\Sigma\,(y-\bar{y})^2}$$

Multiple Linear Regression (MLR)

- In any experiment, there are two types of variables. One variable which is an output variable and there are some variables which are causing that output. When we try to find out the relationship between the input and output variables that will be called as a **Model**.
- Model is simply the **Representation of a Relationship.**
- For example if y is **yield of a crop**, the yield of a crop depends on several factors like **quantity of fertilizers, , irrigation, rainfall, temperature** and so on.
- Yield- y, Quantity of fertilizer- x1(kg), Irrigation level -x2(cm), temperature- x3(^{0}C)

y is dependent variable while x1, x2 and x3 are independent variables.

y=2x1+3x2+4x3 As more than one independent variable is involved it is called **Multiple Linear Regression**

- For example relief to patient depends on quantity of dosage, BP of patients, Sugar level of patients.

$y=4x1+2x2+3x3$ This is also example of **Multiple Linear Regression (MLR)**

- There is some mathematical relationship which is existing in the nature.
- The problem is that **we don't know that relationship.**
- By knowing this relationship, it will help us for the **better future and better planning.**

A model is good when it **incorporates all salient features** of the phenomenon.

Multiple Linear Regression (MLR) Derivation

- For univariate or simple linear regression $y = \beta_0 + \beta_1 x + \varepsilon$
- For multivariate or multiple linear regression $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_k x_k + \varepsilon$

β_0 : **Intercept Term** $\beta_1, \beta_2, \beta_3, \dots \beta_k$ **: Regression Coefficients**

ε : **Random error** Here random error is associated with every variable i.e. $(x_1, x_2, x_3 \dots x_k)$

We will get vector for random error

$\hat{y} = \beta_0 + \beta_1 x$ $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_k x_k$

- If the number of explanatory variables becomes very large then situation becomes more critical.
- We want to retain only the important variable which are trying to affect the outcome y

Model Setup:

Conducted experiment n time

Yield- y,

Quantity of fertilizer- x_1(kg),

Irrigation level -x_2(cm),

Temperature- x_3(^{0}C),

Rain-x_4(mm)

y	x_1	x_2	x_3	x_4	....	x_k
y_1	x_{11}	x_{12}	x_{13}	x_{14}	....	x_{1k}
y_2	x_{11}	x_{12}	x_{13}	x_{14}	....	x_{1k}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
y_n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	....	x_{nk}

Equation of the Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_k x_k + \varepsilon$$

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} \ldots \beta_k x_{1k} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} \ldots \beta_k x_{2k} + \varepsilon_2$$

$$y_3 = \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} \ldots \beta_k x_{3k} + \varepsilon_3$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} \ldots \beta_k x_{nk} + \varepsilon_n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon n \end{bmatrix}$$

Dimensions	Y	X	β	ε

Y=n×1

X=n×(k+1)

β=(k+1)×1

ε=n×1

Actual Equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots . \beta_k x_k + \varepsilon$

Predicted Equation $\hat{y} = \beta_0 + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_3 \dots . \widehat{\beta_k} x_k$

$$\boxed{\widehat{\beta} = (X'X)^{-1} X'Y}$$

Derivation of MLR by Least Square Method

Aim is to find $\beta_0, \beta_1, \beta_2, \beta_3 \ldots\ldots\ldots\ldots\ldots\ldots\ldots \beta_k$

That minimizes $\sum_{i=1}^{n} \epsilon_i^2$

Actual Equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots. \beta_k x_k + \varepsilon$

Predicted Equation $\hat{y} = \beta_0 + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_3 \ldots. \widehat{\beta_k} x_k$

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2$$

where $S(\beta)$

Assumption:

- Real values
- Convex
- Differential function

We will get optimized (minimum) value for coefficients

$$S(\beta) = \sum_{i=1}^{n} \varepsilon' \times \varepsilon \qquad Y = X\beta + \varepsilon$$

$$S(\beta) = \sum_{i=1}^{n} \varepsilon' \times \varepsilon$$

$$Y = X\beta + \varepsilon$$

$$\varepsilon = Y - X\beta$$

$$S(\beta) = (Y - X\beta)' \times (Y - X\beta)$$

$$S(\beta) = (Y' - \beta' X') \times (Y - X\beta)$$

$$S(\beta) = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta$$

$$S(\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

In order to find the minimum of the sum of squares, we take the gradient with respect to

and set it equal to zero.

$$S(\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

$$\frac{d(S(\beta))}{d\beta} = 0 \qquad \frac{d(Y'Y - 2Y'X\beta + \beta'X'X\beta)}{d\beta} = 0$$

$$-2X'Y\beta + \boxed{\beta'X'X\beta} = 0$$

$$\frac{d(\beta'A\beta)}{d\beta} = 2A\beta \qquad A = XX' \qquad 2X'X\beta - 2X'Y = 0$$

$$2X'X\beta - 2X'Y = 0$$

$$(X'X)\beta = X'Y$$

$$(X'X)(X'X)^{-1}\beta = X'Y(X'X)^{-1}$$

$$\beta = X'Y(X'X)^{-1}$$

$$\beta = (X'X)^{-1}X'Y$$

$$\boxed{\widehat{\beta} = (X'X)^{-1}X'Y}$$

Final Equation of MLR

Logistic Regression: What And Why?

Logistic Regression produces results in a binary format which is used to predict the outcome of a categorical dependent variable. So the outcome should be discrete/ categorical such as:

Linear vs Logistic Regression

Basis	Linear Regression	Logistic Regression
Core Concept	The data is modelled using a straight line	The probability of some obtained event is represented as a linear function of a combination of predictor variables.
Used with	Continuous Variable	Categorical Variable
Output/Prediction	Value of the variable	Probability of occurrence of event
Accuracy and Goodness of fit	measured by loss, R squared, Adjusted R squared etc.	Accuracy, Precision, Recall, F1 score, ROC curve, Confusion Matrix, etc

Logistic Regression Equation

The Logistic Regression Equation is derived from the Straight Line Equation

Equation of a straight line

$Y = C + BIX1 + B2X2 + \ldots.$ → Range is from –(infinity) to (infinity)

Let's try to reduce the Logistic Regression Equation from Straight Line Equation

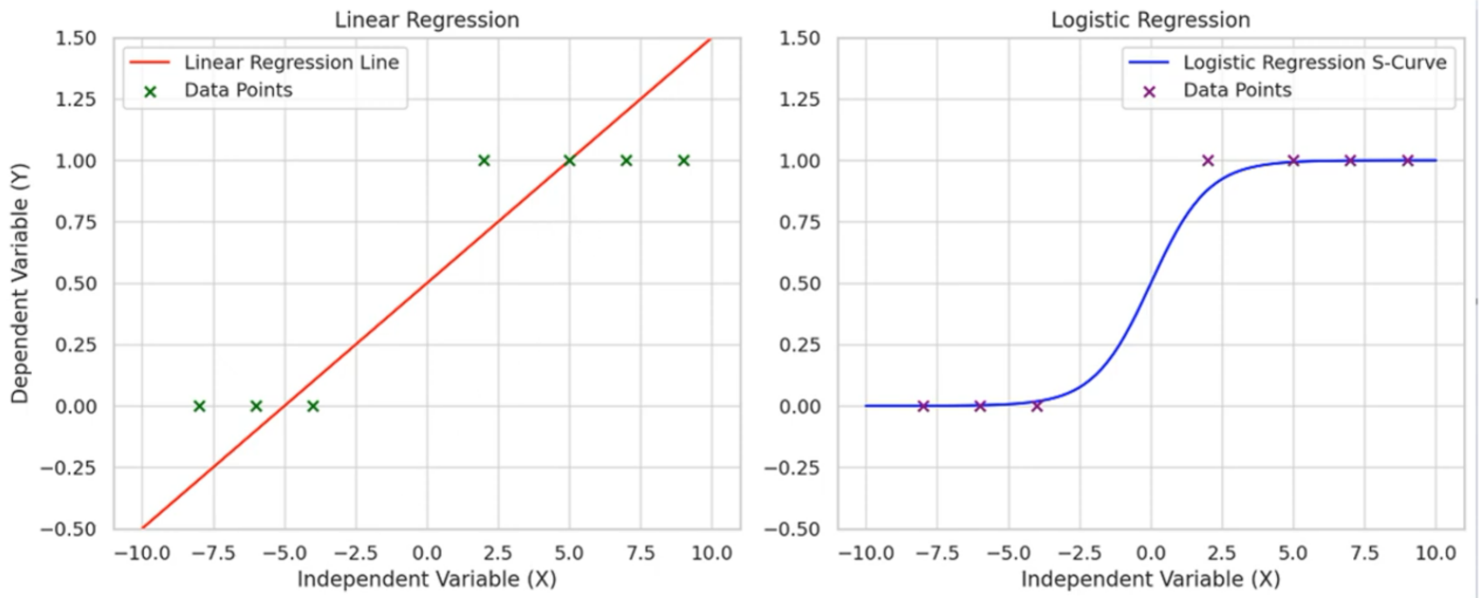
$Y = C + BIX1 + B2X2 + \ldots.$ → In Logistic equation Y can be only from 0 to 1

Now , to get the range of Y between 0 and infinity, let's transform Y

$\frac{Y}{1-Y}$ Y= 0 then 0
Y= 1 then infinity → Now, the range is between 0 to infinity

Let us transform it further, to get range between –(infinity) and (infinity)

$\log\left[\frac{Y}{1-Y}\right] \Rightarrow Y = C + BIX1 + B2X2 + \ldots.$ → **Final Logistic Regression Equation**



Linear Regression: Shows a straight line which can produce outputs beyond the range of 0 and 1, making it unsuitable for classification.

Logistic Regression: Shows the S-shaped curve which restricts the output between 0 and 1, making it ideal

- The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \exp \; erience \;)}}$$

- where $z=\beta_0+\beta_1x$ for a single feature. β_0 is the intercept, and β_1 is the coefficient for the feature x.
- As z approaches **positive ∞, σ(z) approaches 1, and as z approaches negative ∞, σ(z) approaches 0.**
- This “S” shape ensures that the probability output never exceeds the range of 0 to 1, which suits binary outcomes like "yes" or "no".

Logistic Regression Equation

$$\frac{p(x)}{1-p(x)} = e^z$$

Applying natural log on odd. then log odd will be:

$$\log\left[\frac{p(x)}{1-p(x)}\right] = z$$

$$\log\left[\frac{p(x)}{1-p(x)}\right] = w \cdot X + b$$

$$\frac{p(x)}{1-p(x)} = e^{w \cdot X+b} \quad \cdots \text{Exponentiate both sides}$$

$$p(x) = e^{w \cdot X+b} \cdot (1 - p(x))$$

$$p(x) = e^{w \cdot X+b} - e^{w \cdot X+b} \cdot p(x))$$

$$p(x) + e^{w \cdot X+b} \cdot p(x)) = e^{w \cdot X+b}$$

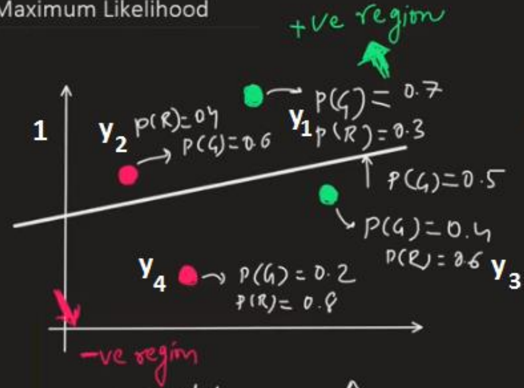
$$p(x)(1 + e^{w \cdot X+b}) = e^{w \cdot X+b}$$

$$p(x) = \frac{e^{w \cdot X+b}}{1 + e^{w \cdot X+b}}$$

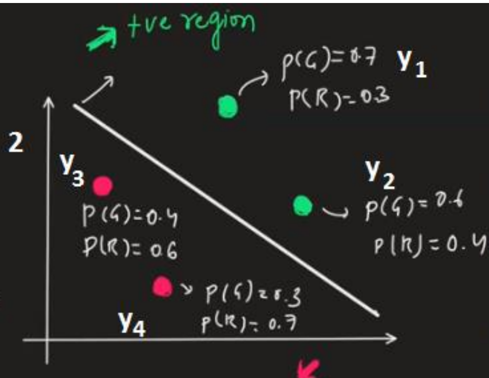
then the final logistic regression equation will be:

$$p(X; b, w) = \frac{e^{w \cdot X+b}}{1+e^{w \cdot X+b}} = \frac{1}{1+e^{-w \cdot X+b}}$$

Maximum Likelihood



$\hat{y}_i$



1

$\hat{y}_1$	0.7	$1-\hat{y}_1$	0.3
$\hat{y}_2$	0.6	$1-\hat{y}_2$	0.4
$\hat{y}_3$	0.4	$1-\hat{y}_3$	0.6
$\hat{y}_4$	0.2	$1-\hat{y}_4$	0.8

2

$\hat{y}_1$	0.7	$1-\hat{y}_1$	0.3
$\hat{y}_2$	0.6	$1-\hat{y}_2$	0.4
$\hat{y}_3$	0.4	$1-\hat{y}_3$	0.6
$\hat{y}_4$	0.3	$1-\hat{y}_4$	0.7

1

Y_1	1
Y_2	0
Y_3	1
Y_4	0

2

Y_1	1
Y_2	1
Y_3	0
Y_4	0

model 1 $\rightarrow$ $0.7 \times 0.4 \times 0.4 \times 0.8 = 0.089$

model 2 $\rightarrow$ $0.7 \times 0.6 \times 0.6 \times 0.7 = 0.176$

$\log(ab) = \log a + \log b$

$\log(\max) = -\log(0.7) - \log(0.4) - \log(0.4) - \log(0.8)$

$0-1 = -ve$

cross entropy minimize

$\log(0.1) > \log(0.9)$

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

MSE

$$\boxed{L = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)}$$

closed form

gradient descent

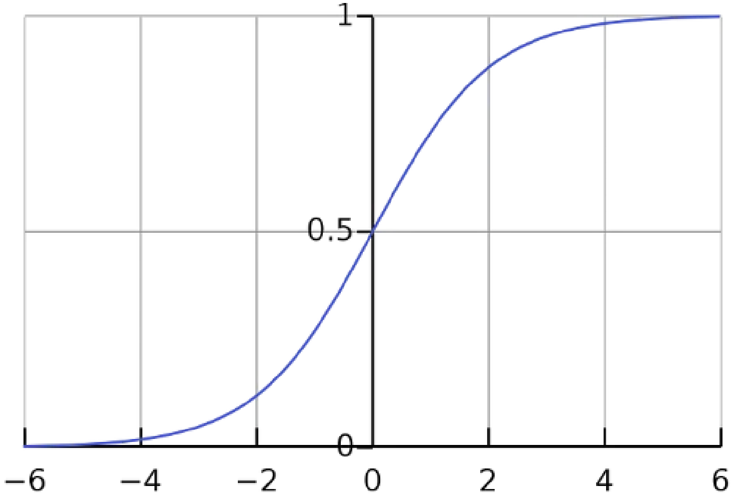
└→ log-loss error

└→ Binary cross entropy

min w_1, w_2, w_0

Sigmoid Function

$$g(x) = \frac{1}{1 + e^{-x}}$$



We can see that its upper bound is 1 and lower bound is 0, this property makes sure we output a probability.

Derivative of sigmoid:

$$g'(x) = \frac{(1)'e^{-x} - 1(e^{-x})'}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})} \frac{e^{-x}}{(1+e^{-x})}$$

$$= \frac{1}{(1+e^{-x})} \frac{1+e^{-x}-1}{(1+e^{-x})} = \frac{1}{(1+e^{-x})} \left(1 - \frac{1}{(1+e^{-x})}\right)$$

$$==> g'(x) = g(x)(1 - g(x))$$

Likelihood Function for Logistic Regression

The predicted probabilities will be:

- for y=1 The predicted probabilities will be: p(X;b,w) = p(x)
- for y = 0 The predicted probabilities will be: 1-p(X;b,w) = 1-p(x)

Suppose we have a matrix of features and a vector of corresponding targets:

$$\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}) \in R^{N,D}$$

$$\mathbf{y} = (y_1, y_2, ..., y_n) \in R^N,\ y_i \in \{0, 1\}$$

where N is number of data points and D is number of dimension at each data point.

Linear transformation **h** mapping from **X** to **y** by parameter **w**:

$$\mathbf{h} = \mathbf{X}\mathbf{w} \in R^N$$

Apply element-wise of sigmoid function **z** to **h** :

$$\mathbf{z} = \sigma(\mathbf{h}) = P(y = 1|x) = \frac{1}{1 + e^{-\mathbf{h}}}$$

Since sigmoid outputs probability, we use negative log likelihood to represent the error:

$$J = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(z_i) + (1 - y_i)log(1 - z_i))$$

where N is number of data points, yi is true label, zi is predicted probability of sigmoid. We want to minize this loss with respect to parameters **w**.

$$\text{Use chain rule: } \frac{\partial J}{\partial \mathbf{w}} = \frac{\partial J}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{w}}$$

$$\frac{\partial J}{\partial \mathbf{z}} = -\frac{1}{N}\left(\frac{\mathbf{y}}{\mathbf{z}} - \frac{1-\mathbf{y}}{1-\mathbf{z}}\right) = \frac{1}{N}\left[\frac{\mathbf{z}-\mathbf{y}}{\mathbf{z}(1-\mathbf{z})}\right]$$

$$\frac{\partial \mathbf{z}}{\partial \mathrm{h}} = \mathbf{z}(1-\mathbf{z})$$

$$\frac{\partial \mathrm{h}}{\partial \mathbf{w}} = \mathbf{X}$$

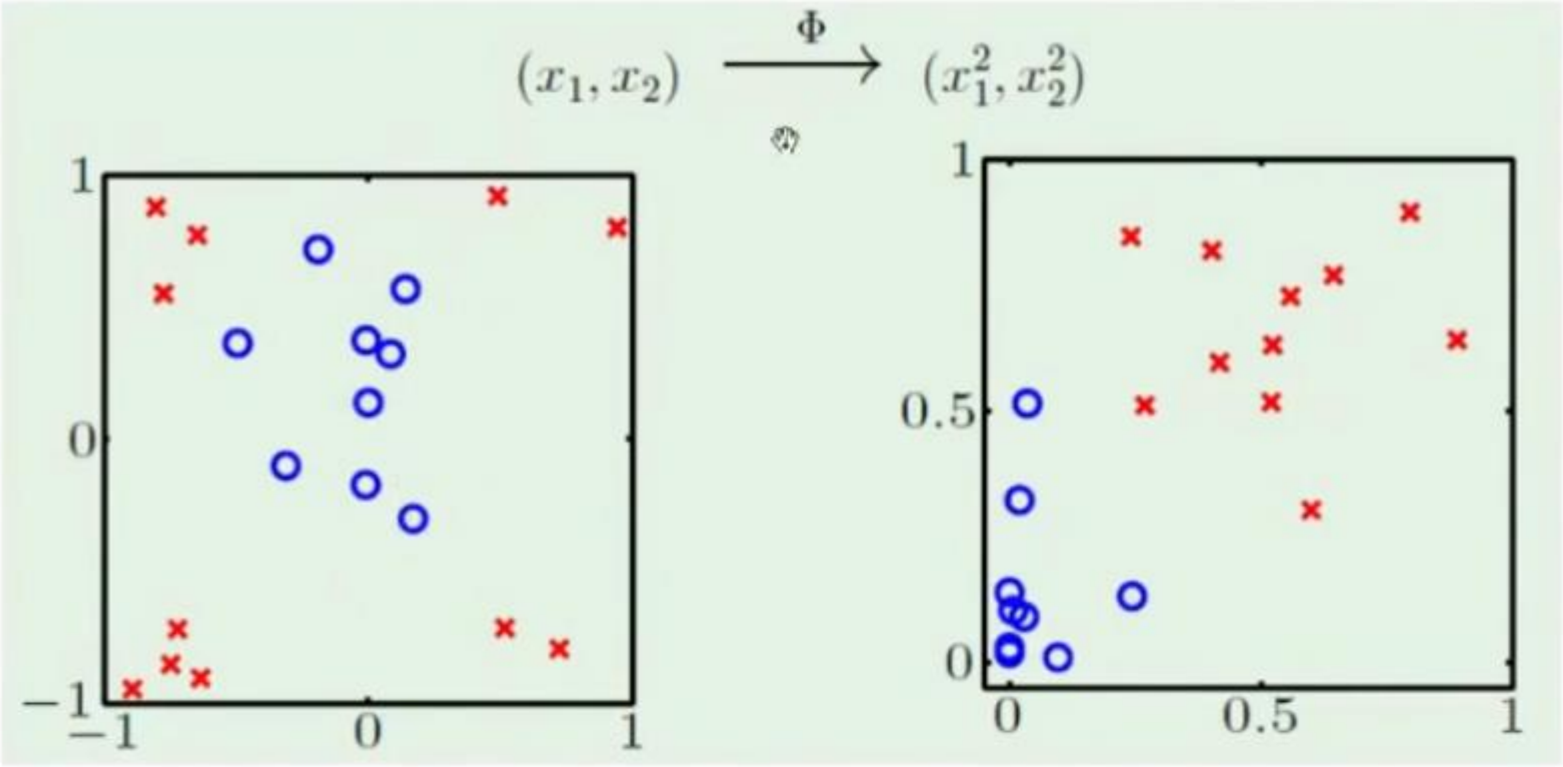
$$\frac{\partial J}{\partial \mathbf{w}} = \frac{1}{N}\left[\mathbf{X}^T(\mathbf{z}-\mathbf{y})\right]$$

Surprisingly, the derivative **J** with respect to **w** of logistic regression is identical with the derivative of linear regression. The only difference is that the output of linear regression is **h** which is linear function, and in logistic is **z** which is sigmoid function.

After found derivative we use gradient descent to update the parameters:

$$\text{Gradient descent: } \mathbf{w} = \mathbf{w} - \alpha\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \text{ with } \alpha \text{ is the learning rate hyperparameter}$$

Non-linear Regression



What is non-linear regression?

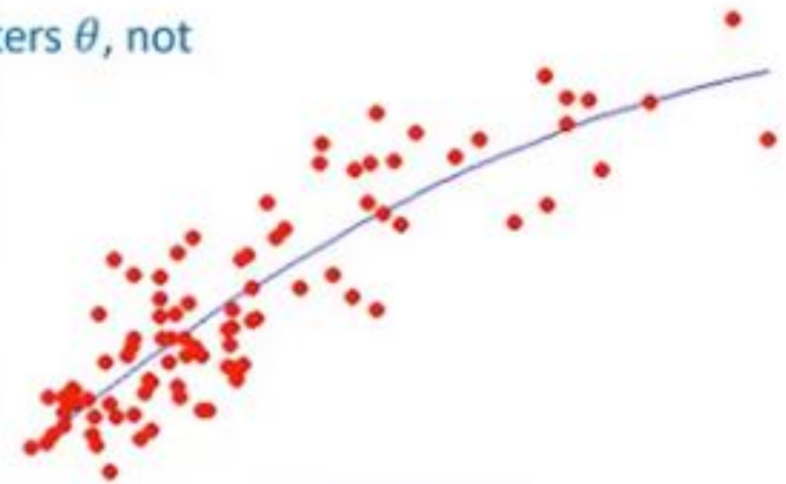
- To model non-linear relationship between the dependent variable and a set of independent variables
- $\hat{y}$ must be a non-linear function of the parameters θ, not necessarily the features x

$$\hat{y} = \theta_0 + {\theta_2}^2 x$$

$$\hat{y} = \theta_0 + \theta_1 {\theta_2}^x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + {\theta_1}^{(x-\theta_2)}}$$



Linear vs non-linear regression

- How can I know if a problem is linear or non-linear in an easy way?
  - Inspect visually
  - Based on accuracy

- How should I model my data, if it displays non-linear on a scatter plot?
  - Polynomial regression
  - Transform your data

What is polynomial regression?

- Some curvy data can be modeled by a polynomial regression
- For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

- A polynomial regression model can be transformed into linear regression model.

$$x_1 = x$$
$$x_2 = x^2$$
$$x_3 = x^3$$

$$\boxed{\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3}$$