Artificial Intelligence

Dr. Mubashir Ahmad (Ph.D.)

Regression Model

- Regression modeling is a process of determining a relationship between one or more independent variables and one dependent or output variable.
- **Examples:**
  1. Predicting the height of a person given the age of the person.
  2. Predicting the price of the car given the car model, year of manufacturing, mileage, engine capacity, etc.

Types Regression Models

- Based on the type of functions used to represent the relationship between the dependent or output variable and independent variables, the regression models are categorized into four types. The regression models are,
  1. Simple Linear Regression
  2. Multiple Regression
  3. Polynomial Regression
  4. Logistic Regression

Simple Linear Regression

1. Simple Linear Regression

- Assume that there is only one independent variable x. If the relationship between x (independent variable) and y (dependent or output variable) is modeled by the relation,

$$y = a + bx$$

- then the regression model is called a linear regression model.

Multiple Regression

2. Multiple Regression

- Assume that there are multiple independent variables *say* x_1, x_2, ...x_n. If the relationship between independent variables x and dependent or output variable y is modeled by the relation,

$$y=a_0+a_1*x_1+a_2*x_2+.....+a_n*x_n$$

- then the regression model is called a multiple regression model.

Polynomial Regression

3. Polynomial regression

- Assume that there is only one independent variable x. If the relationship between independent variables x and dependent or output variable y is modeled by the relation,

$$y=a_0+a_1*x+a_2*x^2+\dots\dots\dots+a_n*x^n$$

- for some positive integer n>1, then we have a polynomial regression.

Logistic Regression

4. Logistic Regression

- Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature.

Solved Example 1 Simple linear regression model

- Apply a simple linear regression model to get the simple linear regression equation. If the equation of the simple linear regression model is given below
- $Y = a + bX$
- First we need to find the a and b, for this purpose first find the mean and covariance.
- Before doing this we are looking at the dataset which is given below.

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

Solved Example 1 Simple linear regression model

- X bar is the mean of all the data points. And n is the data points.
- Y bar is the mean of Y and n is the number of data points in this case.

Means of x and y are given by

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

Solved Example 1 Simple linear regression model

- n is the data points. xi is the data point and xi bar is the mean of x.
- yi is the data point and yi bar is the mean of x.

Variance of x is given by,

$$\mathrm{Var}\,(x) = \frac{1}{n-1}\sum(x_i - \bar{x}_i)^2$$

The covariance of x and y, denoted by Cov(x, y)is defined as

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

$$b = \frac{\mathrm{Cov}\,(x, y)}{\mathrm{Var}\,(x)}$$

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

Solved Example 1 Simple linear regression model

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

$$n = 5$$

$$\bar{x} = \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0)$$

$$= 3.0$$

$$\bar{y} = \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25)$$

$$= 2.06$$

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

$$\mathrm{Cov}\,(x, y) = \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)]$$

$$= 1.0625$$

Solved Example 1 Simple linear regression model

$$\mathrm{Var}(x) = \frac{1}{n-1}\sum(x_i - \bar{x}_i)^2$$

$$\mathrm{Var}(x) = \frac{1}{4}[(1.0-3.0)^2 + \cdots + (5.0-3.0)^2]$$
$$= 2.5$$

$$b = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

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

$$b = \frac{1.0625}{2.5}$$
$$= 0.425$$
$$a = 2.06 - 0.425 \times 3.0$$
$$= 0.785$$

Therefore, the linear regression model for the data is

y = a + bx

if $x = 9$

$y = 0.785 + 0.425 \times 9$

$\Longrightarrow$ ——

Solved Example 1 Simple linear regression model

X	Y	Y^ = 0.785 + 0.425 * X	$(Y - Y^{\wedge})^2$
1	1	1.21	0.0441
2	2	1.63	0.1369
3	1.3	1.3375	2.76
4	3.75	2.3787	1.88
5	2.25	1.741	0.26
SUM			5.07

Formulas for error

1. **Mean Absolute Error (MAE):**
   $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
   where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

2. **Mean Squared Error (MSE):**
   $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

3. **Mean Absolute Percentage Error (MAPE):**
   $\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$

4. **Mean Percentage Error (MPE):**
   $\mathrm{MPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i}$

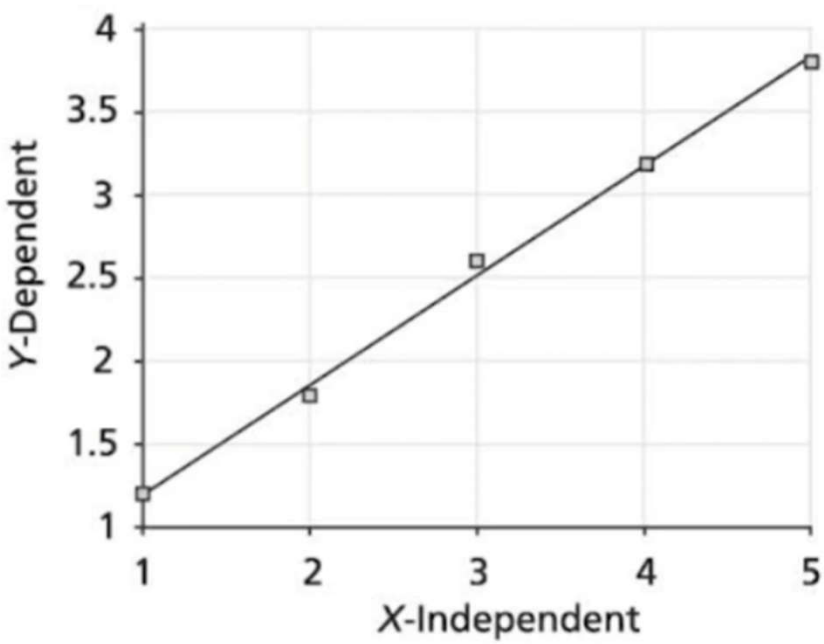
5. **Root Mean Squared Error (RMSE):**
   $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Solved Example 2 Simple Linear Regression

- Let us consider an example where the five weeks' sales data (in Thousands) is given as shown in Table.
- Apply linear regression technique to predict the 7th and 12th week sales.

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Solved Example 2 Simple Linear Regression



x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Solved Example 2 Simple Linear Regression

- Linear regression equation is given by

$$y = a_0 + a_1 * x + e$$

$$where$$

$$a_1 = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - \bar{x}^2}$$

- a0 is the constant, e is the error, and a1 is the coefficient of the independent variable.
- To calculate the mean of XY, X, and Y.
- Then calculate x^2 and its mean, and then the mean of x and its square.

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

$$a_0 = \bar{y} - a_1 * \bar{x}$$

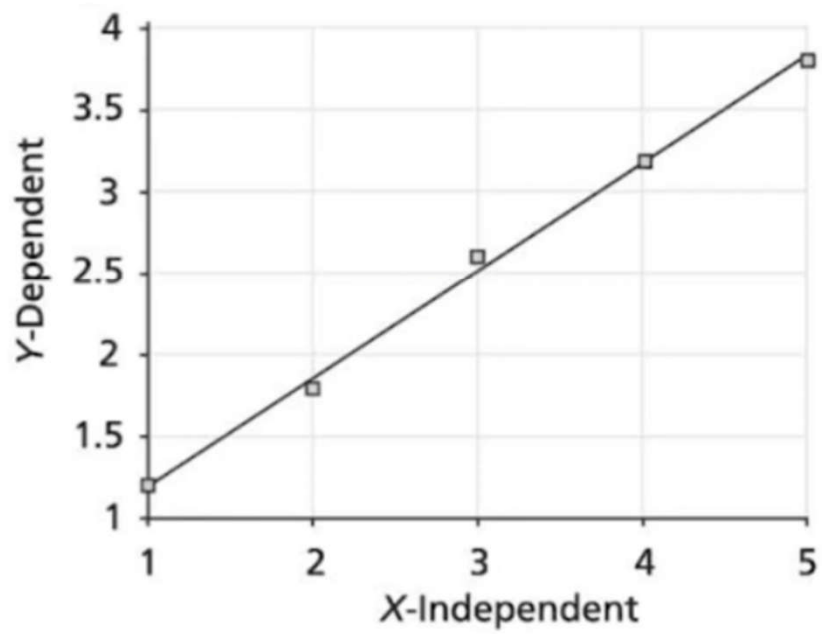
Solved Example 2 Simple Linear Regression

	x_i (Week)	y_j (Sales in Thousands)	x_i^2	$x_i * y_j$
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
Sum	15	12.6	55	44.4
Average	$\overline{x} = 3$	$\overline{y} = 2.52$	$\overline{x^2} = 11$	$\overline{xy} = 8.88$

Solved Example 2 Simple Linear Regression

- $\bar{x} = 3$ $\bar{y} = 2.52$ $\overline{x^2} = 11$ $\overline{xy} = 8.88$

- $a_1 = \frac{(\overline{xy})-(\bar{x})(\bar{y})}{\overline{x^2}-\bar{x}^2} = \frac{8.88-3*2.52}{11-3^2} = 0.66$
- $a_0 = \bar{y} - a_1 * \bar{x} = 2.52 - 0.66 * 3 = 0.54$
- **Regression equation is**
- $\boldsymbol{y = a_0 + a_1 * x}$
- $\boldsymbol{y = 0.54 + 0.66 * x}$



Solved Example 2 Simple Linear Regression

Regression equation is

$$y = a_0 + a_1 * x$$

$$y = 0.54 + 0.66 * x$$

The predicted 7th week sale (when x = 7) is,

y = 0.54 + 0.66 x 7 = 5.16

the predicted 12th week sale (when x = 12) is,

y = 0.54 + 0.66 x 12 = 8.46

Solution 2 of the previous question

x_i (Week)	y_j (Sales in Thousands)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

$$a_0 = \bar{y} - a_1 * \bar{x}$$

$y = a_0 + a_1 * x + e$

$\Rightarrow \quad y = a_0 + a_1 x$

$y = a + bx$

(Solution 2)

$\Rightarrow \quad \bar{x}_i = 3$

$\bar{y}_i = 2.52$

$$Cov(x,y) = \frac{1}{4}\Big[(1-3)(1.2-2.52) + (2-3)(1.8-2.52) + (3-3)(2.6-2.52) + (4-3)(3.2-2.52) + (5-3)(3.8-2.52)\Big]$$

$$\Rightarrow \frac{1}{4}\Big[(-2)(-1.32) + (-1)(-0.72) + (0)(0.08) + (1)(0.68) + (2)(1.28)\Big]$$

$$\Rightarrow \frac{1}{4}\Big[(2.64) + (0.72) + 0 + 0.68 + 2.56\Big]$$

$$\Rightarrow \frac{1}{4}\Big[6.6\Big] \Rightarrow 1.65$$

$$Var(x) = \frac{1}{4}\sum (x_i - \bar{x}_i)^2$$

$$\Rightarrow \frac{1}{4}\left[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\right]$$

$$\Rightarrow \frac{1}{4}\left[4 + 1 + 0 + 1 + 4\right]$$

$$\Rightarrow \frac{1}{4}\left[10\right] = \frac{10}{4} = 2.5$$

$$b = a_1 = \frac{Cov(x,y)}{Var(x)} = \frac{1.65}{2.5} = 0.66$$

$$a_0 = \bar{y} - a_1\bar{x} \Rightarrow a_0 = 2.52 - 0.66 * 3$$

$$\Rightarrow 0.54$$

Regression equation

$$y = a_0 + a_1 * x \Rightarrow \boxed{y = 0.54 + 0.61 * x}$$

Solved Example 2 Simple Linear Regression with 2 variables

SUBJECT	Y	X_1	X_2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Solved Example 3 Simple Linear Regression with 2 variables

- Finding the values of *b* (the slopes) is tricky for k>2 independent variables, and you really need matrix algebra to see the computation
- At this point, you should notice that all the terms from the one variable case appear in the two variable case.
- In the two variable case, the other X variable also appears in the equation. For example, X_2 appears in the equation for b_1. s.

Solved Example 3 Simple Linear Regression with 2 variables

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Solved Example 3 Simple Linear Regression with 2 variables

- In our case we need only two independent variables. So equation will be as follows. a is the intercept below. b1 and b2 are the coefficients of x1 and x2.

Y-intercept

Population slopes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Solved Example 3 Simple Linear Regression with 2 variables

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

- We solve the b1 and b2 with the help of these 5 equations.
- Put the values then.

This is not the normal case X1^2, its the variance we need to find.

Var(XY1) is the variance of the product XY1.

E[XY1] is the expected value calculated in step 1.

(XY1 - E[XY1]) is the deviation of XY1 from its mean.

Solved Example 3 Simple Linear Regression with 2 variables

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
$\sum$	19.5	20	24	90	148	99	95.8	45.6

Solved Example 3 Simple Linear Regression with 2 variables

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 10$$

90 − 80

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 32.8$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N} = 17.8$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N} = -48$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 3$$

- Put the values of previously calculated and get the results. After getting these results.

Solved Example 3 Simple Linear Regression with 2 variables

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{32.8 * 17.8 - 3 * (-48)}{10 * 32.8 - 3 * 3} = 2.28$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{10 * (-48) - 3 * 17.8}{10 * 32.8 - 3 * 3} = -1.67$$

Solved Example 3 Simple Linear Regression with 2 variables

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

$$= \frac{19.5}{5} - \frac{2.28 * 20}{5} - \frac{-1.67 * 24}{5} = 2.796$$

- Y bar is the mean of Y and X1 and X2 bar is the mean of X1 and X2.

Solved Example 3 Simple Linear Regression with 2 variables

Final Regression equation or Model is:

$$\boldsymbol{Y = 2.796 + 2.28\, x_1 - 1.67 x_2}$$

$$Now\ given \quad x_1 = 3 \quad and \quad x_2 = 2 \qquad Y = ?$$

$$Y = 2.796 + 2.28 * 3 - 1.67 * 2$$
$$= \mathbf{6.296}$$

Logistic Regression

- Linear regression predicts the numerical response but is not suitable for predicting the categorical variables.
- When categorical variables are involved, it is called classification problem.
- Logistic regression is suitable for binary classification problem.

Logistic Regression

For example, the following scenarios are instances of predicting categorical variables.

1. Is the mail spam or not spam? The answer is yes or no. Thus, categorical dependent variable is a binary response of yes or no.
2. If the student should be admitted or not is based on entrance examination marks. Here, categorical variable response is admitted or not.
3. The student being pass or fail is based on marks secured.

Logistic Regression

How Does the Logistic Regression Algorithm Work?

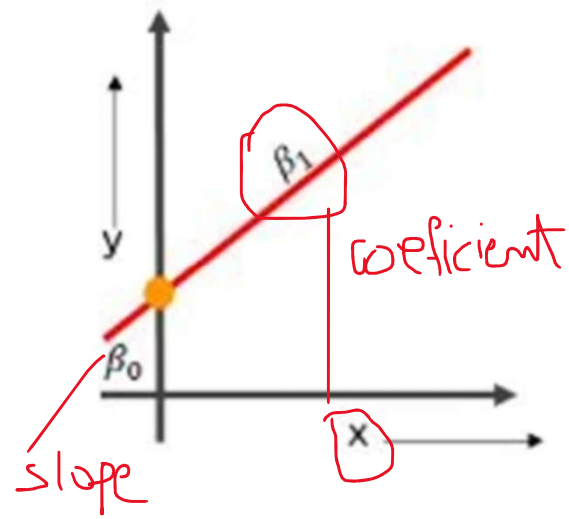
- Consider the following example:
- An organization wants to determine an employee's salary increase based on their performance.
- For this purpose, a linear regression algorithm will help them decide.
- Plotting a regression line by considering the employee's performance as the independent variable, and the salary increase as the dependent variable will make their task easier.

Logistic Regression

- Now, what if the organization wants to know whether an employee would get a promotion or not based on their performance?
- The above linear graph won't be suitable in this case.
- As such, we clip the line at zero and one, and convert it into a sigmoid curve (S curve).
- Based on the threshold values, the organization can decide whether an employee will get a salary increase or not.

Logistic Regression

- To understand logistic regression, let's go over the odds of success.
- $Odds\ (\theta) = \dfrac{Probability\ of\ an\ event\ happening}{Probability\ of\ an\ event\ not\ happening}$
- $Odds\ (\theta) = \dfrac{p}{1-p}$
- The values of odds range from **zero to ∞** and the values of probability lies between **zero and one**.
- Consider the equation of a straight line:
- $\boldsymbol{y = \beta 0 + \beta 1 * x}$



Logistic Regression

- Now to predict the odds of success, we take log on odds formula:

$$\log\left(\frac{p(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x$$

- Exponentiating both the sides, we have:

$$\boxed{e^{\ln(x)} = x}$$

$$e^{ln}\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$$

$$\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$$

Logistic Regression

- $Let\ Y = e^{\beta 0 + \beta 1 * x}$
- $Then \quad \frac{p(x)}{1 - p(x)} = Y$
- $p(x) = Y(1 - p(x))$
- $p(x) = Y - Y(p(x))$
- $p(x) + Y(p(x)) = Y$
- $p(x)(1 + Y) = Y$
- $p(x) = \frac{Y}{1+Y}$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$p(x) = \frac{\frac{\cancel{e^{\beta_0 + \beta_1 x}}}{\cancel{e^{\beta_0 + \beta_1 x}}}}{\frac{1 + e^{\beta_0 + \beta_1 x}}{e^{\beta_0 + \beta_1 x}}}$$

The equation of the sigmoid function is:

$$\boxed{p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}$$

Logistic Regression

The sigmoid curve obtained from the above equation is as follows:

Logistic Regression solved Example 1

- The student dataset has entrance mark based on the historic data of those who are selected or not selected.
- Based on the logistic regression, the values of the learnt parameters are $\beta_0 = 1 \; and \; \beta_1 = 8$.
- Assuming marks of x = 60, compute the resultant class.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$\beta_0 + \beta_1 x = 481$$

$$p(x) = \frac{1}{1 + e^{-481}} = 0.44$$

If we assume the threshold value is 0.5. then it is observed that 0.44 < 0.5, therefore, the candidate with marks 60 is not selected.

Solved example 2 Logistic Regression

Hours Study	Pass (1) / Fail (0)
14	0
15	0
33	1
28	1
39	1

Assume the model suggested by the optimizer for odds of passing the course is,

$$\log(odds) = -64 + 2 * \text{hours}$$

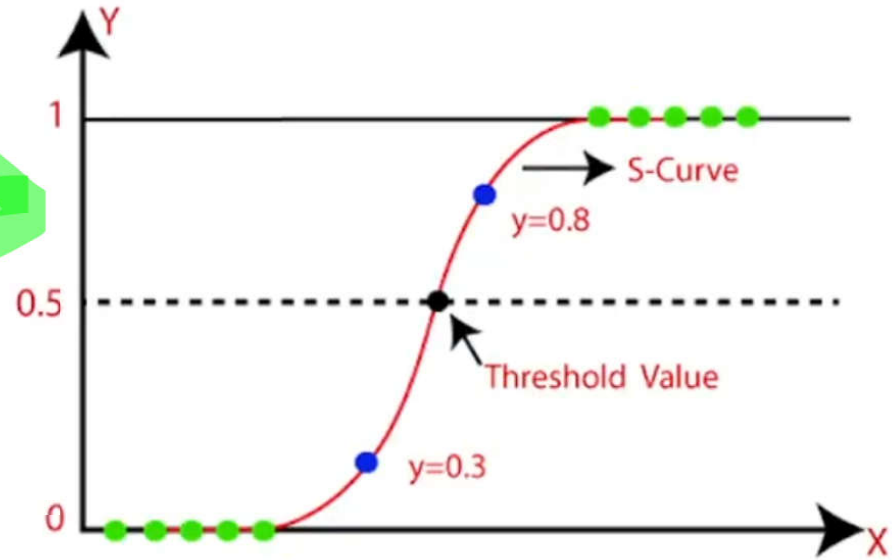
- The dataset of pass or fail in an exam of 5 students is given in the table.
- Use logistic regression as classifier to answer the following questions.

1. Calculate the probability of pass for the student who studied 33 hours.
2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Solved example 2 Logistic Regression

- We use Sigmoid Function in logistic regression

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



Solved example 2 Logistic Regression

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

$$p = \frac{1}{1+e^{-z}}$$

1. Calculate the probability of pass for the student who studied 33 hours.

- We can write the logistic regression function from equation 1 to 2.

Hours Study	Pass (1) / Fail (0)
14	0
15	0
33	1
28	1
39	1

Assume the model suggested by the optimizer for odds of passing the course is,

$$\log(odds) = -64 + 2 * \text{hours}$$

This is assumption

Solved example 2 Logistic Regression

Hours Study	Pass (1) / Fail (0)
14	0
15	0
33	1
28	1
39	1

$$\log(odds) = z = -64 + 2 * hours$$

$$z = -64 + 2 * 33 = -64 + 66 = 2$$

- To get the value of z, we need to put the hours. Then to get the z, we will put it in sigmoid function. P is the probability.

$$p = \frac{1}{1+e^{-2}} = 0.88$$

That is, if student studies 33 hours, then there is **88% chance** that the student will pass the exam

Solved example 2 Logistic Regression

2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
14	0
15	0
33	1
28	1
39	1

$$\boldsymbol{p} = \frac{\mathbf{1}}{\mathbf{1}+e^{-z}} = \mathbf{0.95}$$

$$0.95 * (1 + e^{-z}) = 1$$

$$0.95 * e^{-z} = 1 - 0.95$$

$$e^{-z} = \frac{0.05}{0.95} = 0.0526$$

$$\ln(e^{-z}) = \ln(0.0526)$$

$$\boxed{\ln(e^{x}) = x}$$

$$-z = ln(0.0526) = -2.94$$

$$z = 2.94$$

Solved example 2 Logistic Regression

2. At least how many hours student should study that makes he will pass the course with the probability of more than 95%.

Hours Study	Pass (1) / Fail (0)
14	0
15	0
33	1
28	1
39	1

$z = 2.94$

$\log(odds) = z = -64 + 2 * hours$

$2.94 = -64 + 2 * hours$

$2 * hours = 2.94 + 64$

$2 * hours = 66.94$

$hours = \frac{66.94}{2}$

$\boldsymbol{hours = 33.47\ Hours}$

The student should study **at least 33.47 hours**, so that he will pass the exam with more than 95% probability