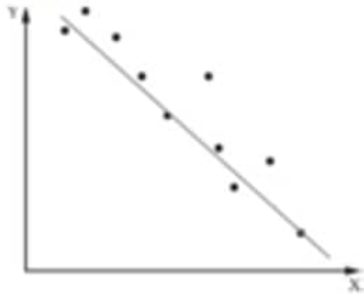
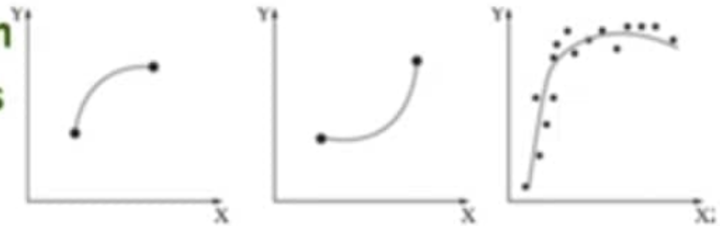
Machine Learning
Subject Code: 20A05602T

UNIT IV

Supervised Learning: Regression – 4-1-1



- Regression,
- **Simple linear regression,**
  - **Slope of the simple linear regression model**
  - **Types of Slopes**
  - **Error in simple regression**
  - **Ordinary Least Squares (OLS) algorithm**
  - **Maximum and minimum point of curves**



Regression

- Regression is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variable(s) (X),
- The dependent variable (Y) is the one whose value is to be predicted.
- This variable is presumed to be functionally related to one (say, X) or more independent variables called predictors.
- to find the function 'f' for the association Y = f (X).

$y = x$

$y = f(x)$

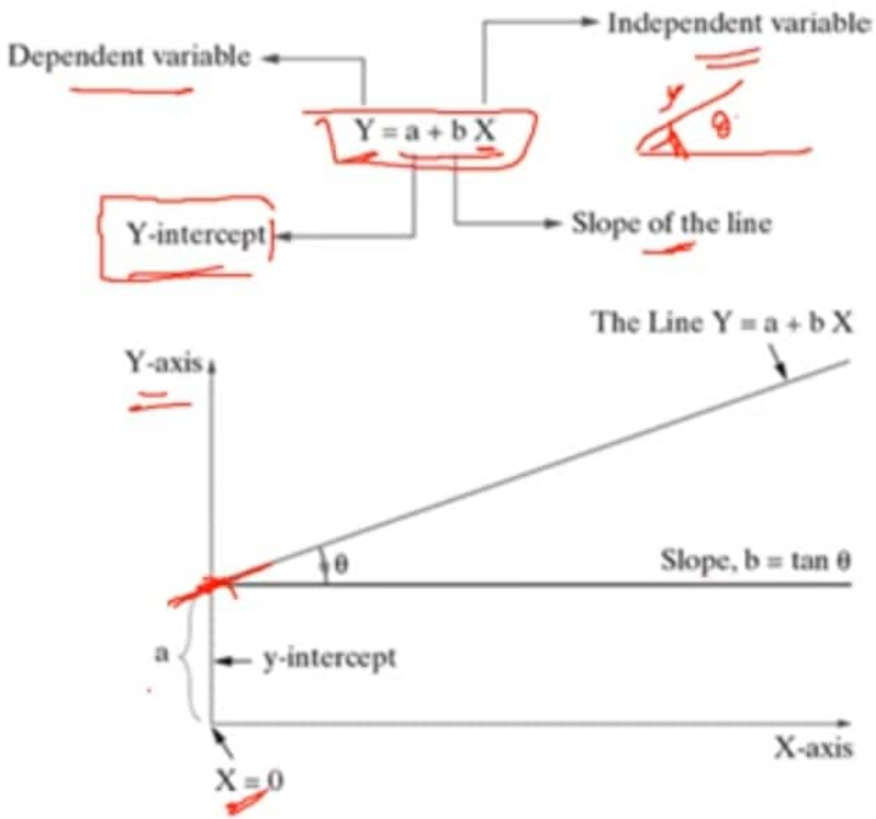
dV

COMMON REGRESSION ALGORITHMS

- The most common regression algorithms are
- Simple linear regression
- Multiple linear regression
- Polynomial regression
- Multivariate adaptive regression splines
- Logistic regression
- Maximum likelihood estimation (least squares)

Simple Linear Regression

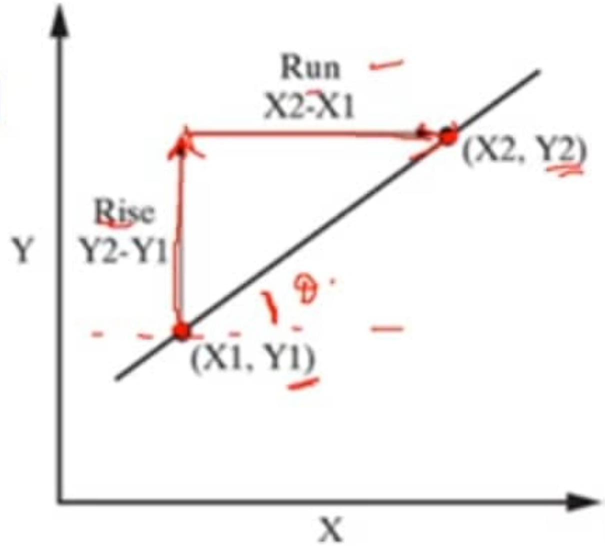
- This model assumes a linear relationship between the dependent variable and the predictor variable
- 'a' is intercept and 'b' is slope of the straight line.
- The value of intercept indicates the value of Y when X = 0.
- it specifies where the straight line crosses the vertical or Y-axis



Slope of the simple linear regression model

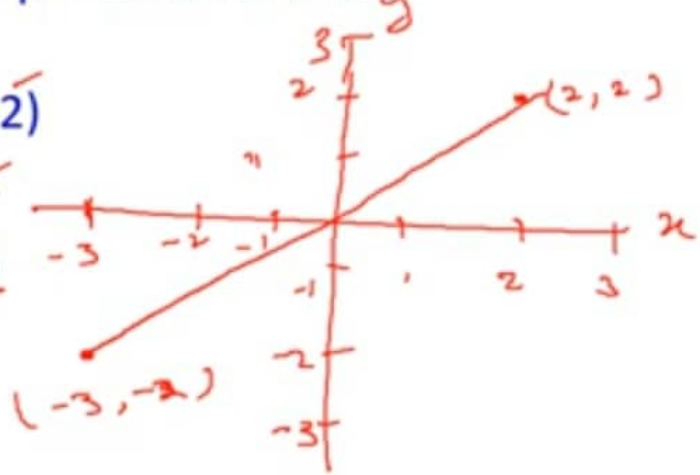
- Slope of a straight line represents how much the line changes in the vertical direction (Y-axis) over a change in the horizontal direction (X-axis).
- Slope = Change in Y/Change in X
- Rise is the change in Y-axis (Y2 – Y1) and
- Run is the change in X-axis (X2 – X1)

$$\text{Slope} = \frac{\text{Rise}}{\text{Run}} = \frac{Y2 - Y1}{X2 - X1}$$



Example

- Let us find the slope of the graph where the lower point on the line is represented as (−3, −2) and the higher point on the line is represented as (2, 2).
- (X1 , Y1) = (−3, −2) and (X2 , Y3) = (2, 2)
- Rise = (Y2 – Y1) = (2 – (−2)) = 2 + 2 = 4
- Run = (X2 – X1) = (2 – (−3)) = 2 + 3 = 5
- Slope = Rise/Run = 4/5 = 0.8

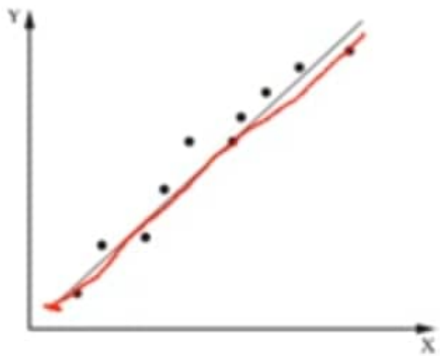


Types of Slopes

- There can be two types of slopes in a linear regression model:
  - positive slope
  - negative slope.
- Different types of regression lines based on the type of slope include
  - Linear Positive Slope
  - Curve Linear Positive Slope
  - Linear Negative Slope
  - Curve Linear Negative Slope

Linear Positive Slope

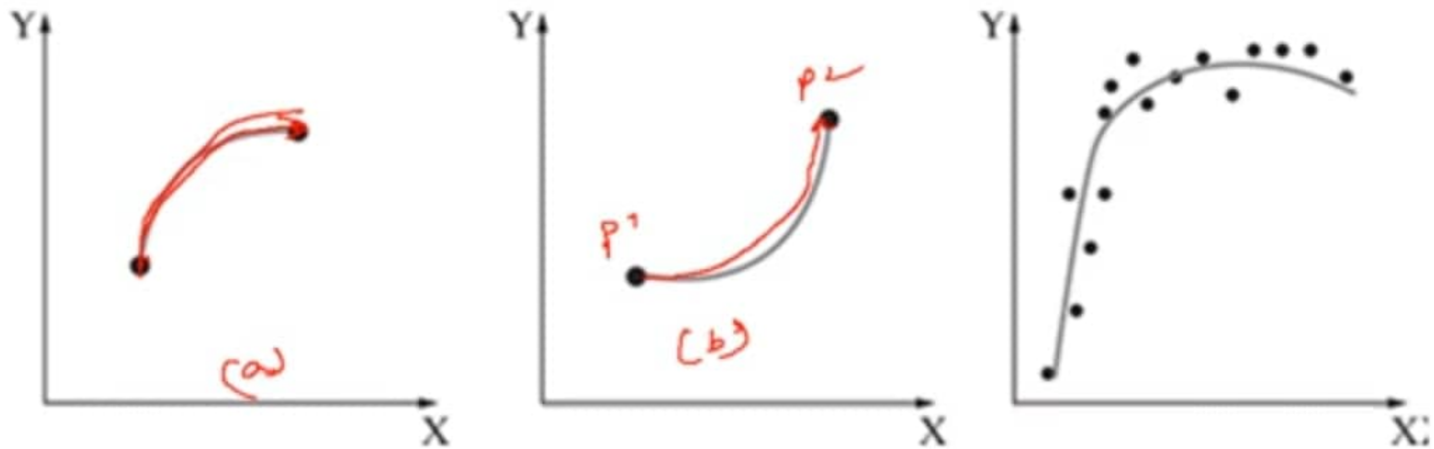
- **Positive slope always moves upward on a graph from left to right**
- Slope = Rise/Run = (Y2 – Y1) / (X2 – X1)
- = Delta (Y) / Delta(X)
- Scenario 1 for positive slope:
- Delta (Y) is positive and Delta (X) is positive
- Scenario 2 for positive slope:
- Delta (Y) is negative and Delta (X) is negative



Subscribe

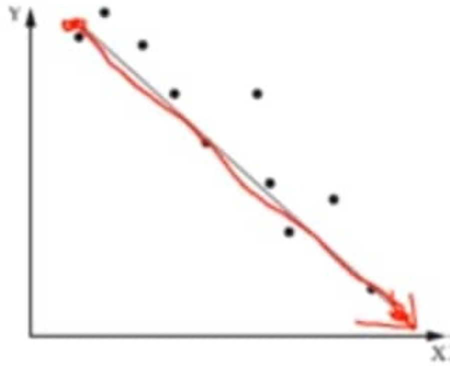
Curve Linear Positive Slope

- **Curves in these graphs slope upward from left to right.**
- Slope = (Y2 – Y1) / (X2 – X1) = Delta (Y) / Delta(X)
- Slope for a variable (X) may vary between two graphs, but it will always be positive; hence, the above graphs are called as graphs with curve linear positive slope.



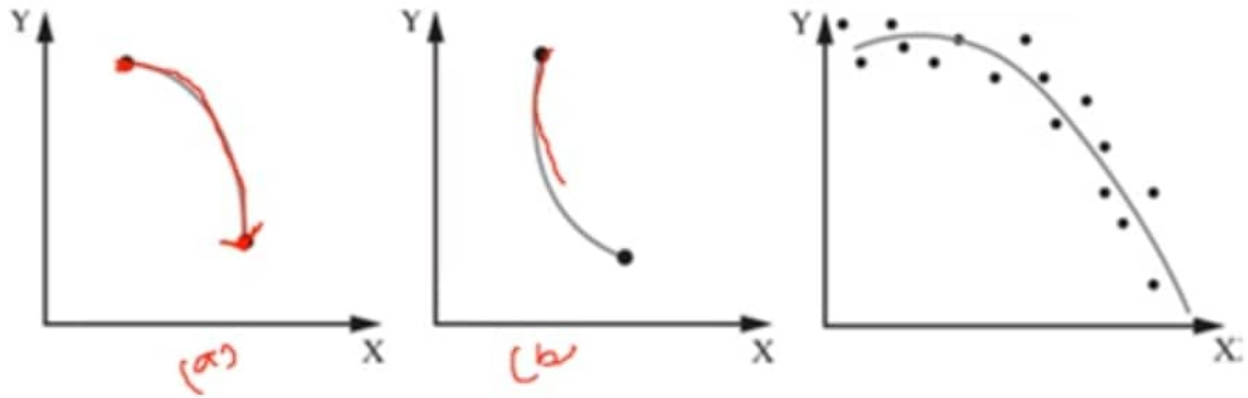
Linear Negative Slope

- A negative slope always moves downward on a graph from left to right. As X value (on X-axis) increases, Y value decreases
- Slope = Rise/Run = (Y2– Y1) / (X2 – X1)
- = Delta (Y) / Delta(X)
- Scenario 1 for negative slope: Delta (Y) is positive and Delta (X) is negative
- Scenario 2 for negative slope: Delta (Y) is negative and Delta (X) is positive



Curve Linear Negative Slope

- Curves in these graphs slope downward from left to right.
- Slope = (Y2 – Y1) / (X2 – X1) = Delta (Y) / Delta(X)
- Slope for a variable (X) may vary between two graphs, but it will always be negative; hence, the above graphs are called as graphs with curve linear negative slope

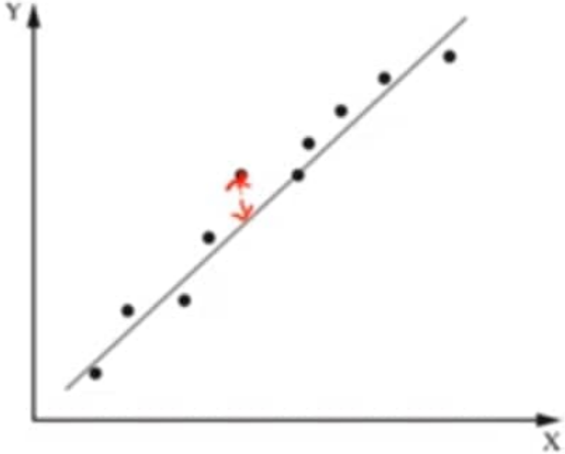


No Relationship Graph

- Scatter graph shown in Figure indicates 'no relationship' curve as it is very difficult to conclude whether the relationship between X and Y is positive or negative.

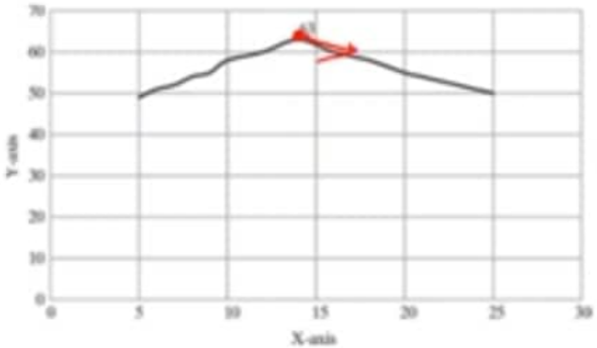
Error in Simple Regression

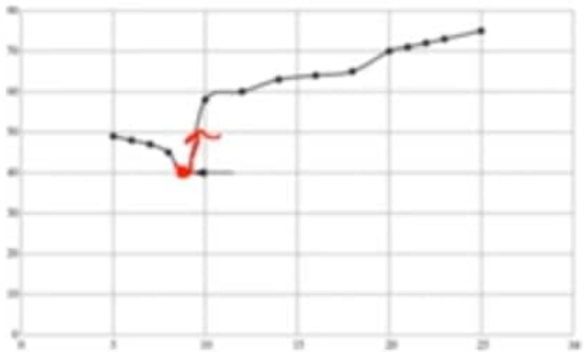
- X and Y values are provided to the machine, and it identifies the values of a (intercept) and b (slope) by relating the values of X and Y.
- identifying the exact match of values for a and b is not always possible.
- There will be some error value (ε) associated with it.
- This error is called marginal or residual error.
- $Y = (a + bX) + \varepsilon$



Maximum and Minimum Point of Curves

- The maximum point is the point on the curve of the graph with the highest y coordinate and a slope of zero. (Point 63 is at the highest point on this curve.)
- The minimum point is the point on the curve of the graph with the lowest y coordinate and a slope of zero. (Point 40 is at the lowest point on this curve.)





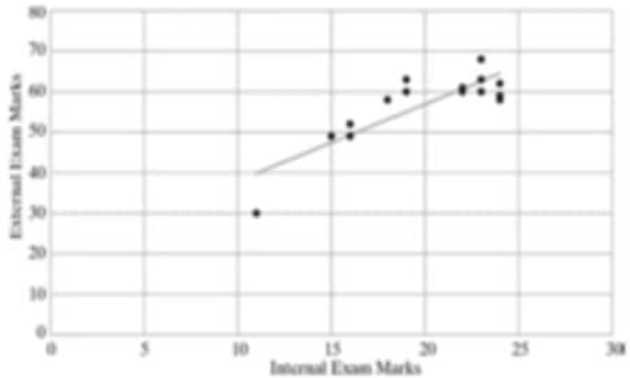
Machine Learning
Subject Code: 20A05602T

UNIT IV - Supervised Learning: Regression – 4-1-2

Simple linear regression,

- Ordinary Least Squares (OLS) algorithm
- Solved Problem
- Residual Error (ε)

Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68



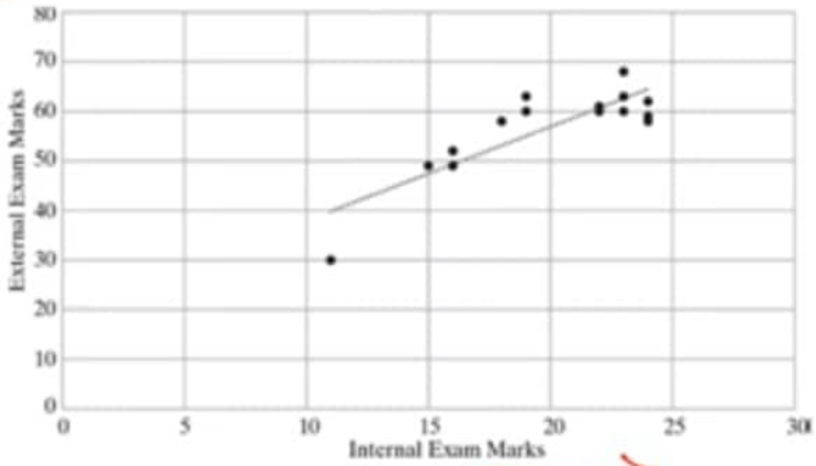
Example of simple regression

- A college professor believes that if the grade for internal examination is high in a class, the grade for external examination will also be high. A random sample of 15 students in that class was selected, and the data is given below:

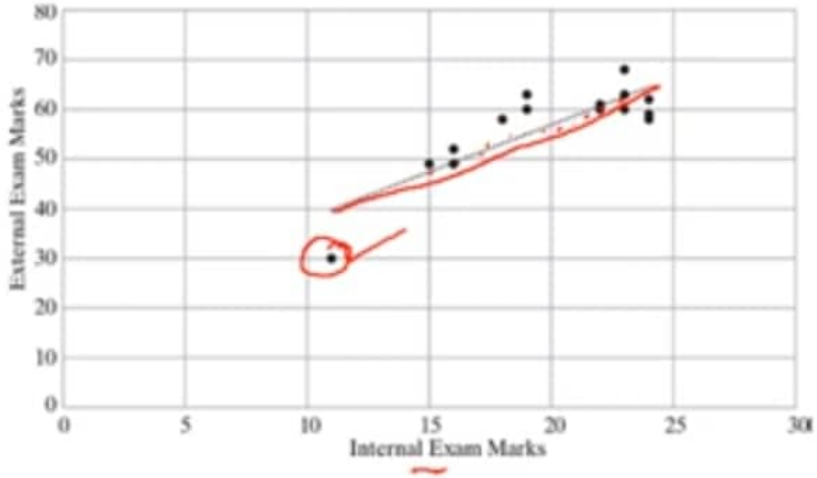
Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

- A scatter plot was drawn to explore the relationship between the independent variable (internal marks) mapped to X-axis and dependent variable (external marks) mapped to Y-axis

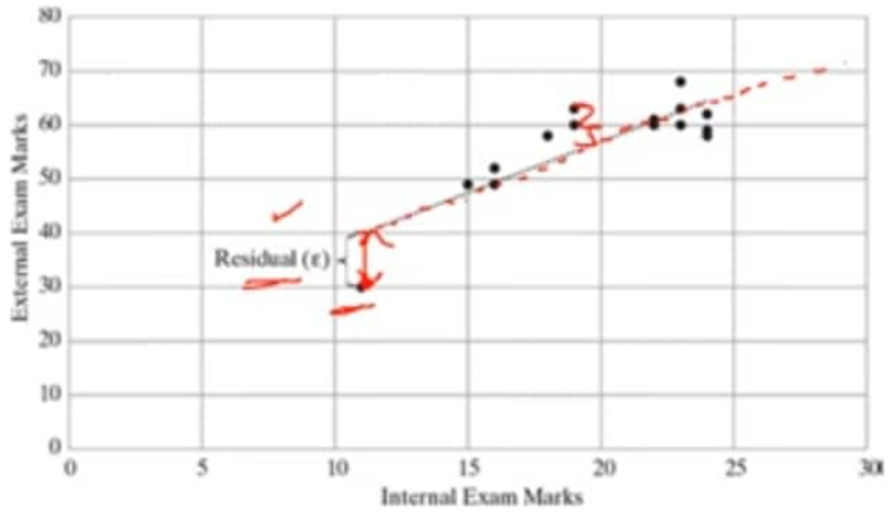


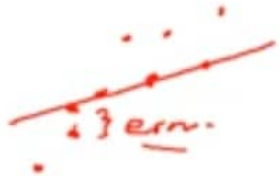
- observe from the graph, the line (i.e. the regression line) does not predict the data exactly. Instead, it just cuts through the data.
- Some predictions are lower than expected, while some others are higher than expected.



- Residual is the distance between the predicted point (on the regression line) and the actual point
- $Y = (a + bX) + \varepsilon$

ε





- Ordinary Least Squares (OLS) is the technique used to estimate a line that will minimize the error (ε),
- which is the difference between the predicted and the actual values of Y.
- This total errors of each prediction or,
- the Sum of the Squares of the Errors (SSE) =

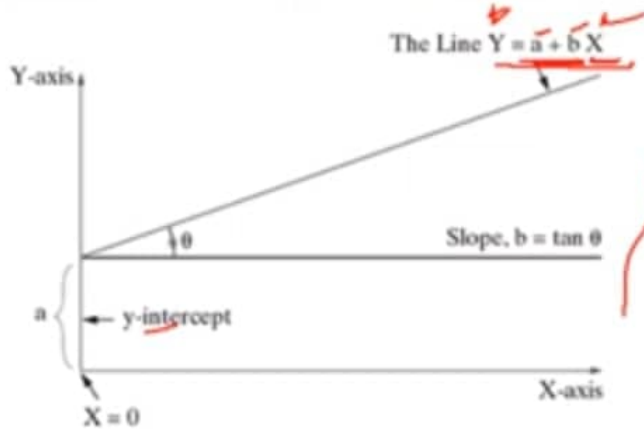
$$\left(\text{i.e. } \sum_i \varepsilon_i^2\right).$$

- It is observed that the SSE is least when b takes the value

$$\boxed{b} = \frac{\Sigma_i (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma_i (X_i - \overline{X})^2} = \frac{\text{cov}(X, Y)}{\text{Var}(X)}$$

- The corresponding value of 'a' calculated using the above value of 'b' is

$$a = \overline{Y} - b\overline{X}$$



Ordinary Least Squares (OLS) algorithm

- Step 1: Calculate the mean of X and Y
- Step 2: Calculate the errors of X and Y
- Step 3: Get the product
- Step 4: Get the summation of the products
- Step 5: Square the difference of X
- Step 6: Get the sum of the squared difference
- Step 7: Divide output of step 4 by output of step 6 to calculate ‘b’
- Step 8: Calculate ‘a’ using the value of ‘b’

Internal Exam (X)	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam (Y)	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

OLS

- Step 1: Calculate the mean of X and Y

$\bar{x}, \bar{y}$

- Step 2: Calculate the errors of X and Y
- Step 3: Get the product
- Step 4: Get the summation of the products
- Step 5: Square the difference of X
- Step 6: Get the sum of the squared difference
- Step 7: Divide output of step 4 by output of step 6 to calculate 'b'
- Step 8: Calculate 'a' using the value of 'b'

$(x - \bar{x})$

Step 2   Step 3   Step 5

X	Y	X- mean (X)	Y- Mean (Y)	$(Xi-\bar{X})(Yi-\bar{Y})$	$(Xi-\bar{X})^2$
15	49	-4.93	-7.8	38.454	24.3049
23	63	3.07	6.2	19.034	9.4249
18	58	-1.93	1.2	-2.316	3.7249
23	60	3.07	3.2	9.824	9.4249
24	58	4.07	1.2	4.884	16.5649
22	61	2.07	4.2	8.694	4.2849
22	60	2.07	3.2	6.624	4.2849
19	63	-0.93	6.2	-5.766	0.8649
19	60	-0.93	3.2	-2.976	0.8649
16	52	-3.93	-4.8	18.864	15.4449
24	62	4.07	5.2	21.164	16.5649
11	30	-8.93	-26.8	239.324	79.7449
24	59	4.07	2.2	8.954	16.5649
16	49	-3.93	-7.8	30.654	15.4449
23	68	3.07	11.2	[illegible]	[illegible]
19.9	56.8		$\Sigma(Xi-\bar{X})(Yi-\bar{Y})$	429.8	226.9335

Step 1   Step 4   Step 6

Step 7: Divide (step4 / step6)

b = 429.28 / 226.93 = 1.89

$b = 1.89$

Step 8: Calculate a using the value of b

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

a = 56.8 – 1.89 × 19.9

a = 19.05

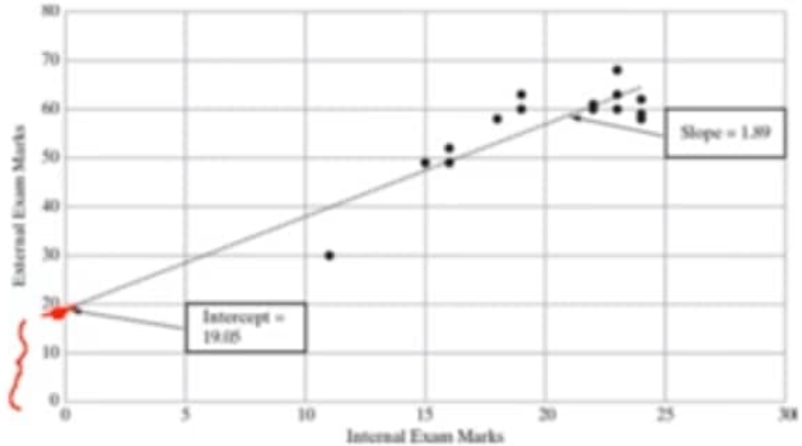
$Y = a + bx$

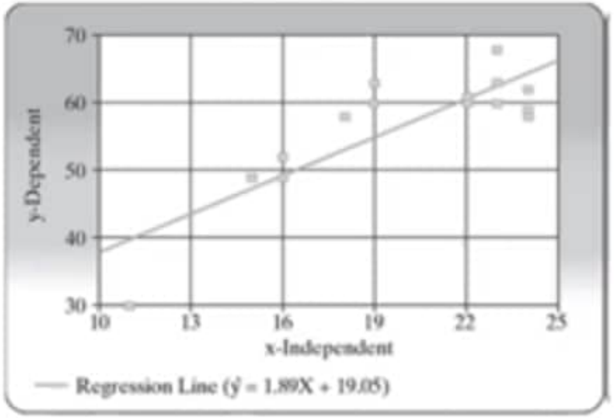
$= 19.05 + 1.89\,X$

Subscribe

- the estimated regression equation is constructed on the basis of the estimated values of a and b:
- $\hat{y} = 1.89395X + 19.0473$
- Marks in external exam = 19.04 + 1.89 × (Marks in internal exam)
- or, MExt = 19.04 + 1.89 × MInt

- The value of the intercept from the above equation is
- 19.05. However, none of the internal mark is 0.
- intercept = 19.05 indicates that 19.05 is the portion of the external examination marks not explained by the internal examination marks.



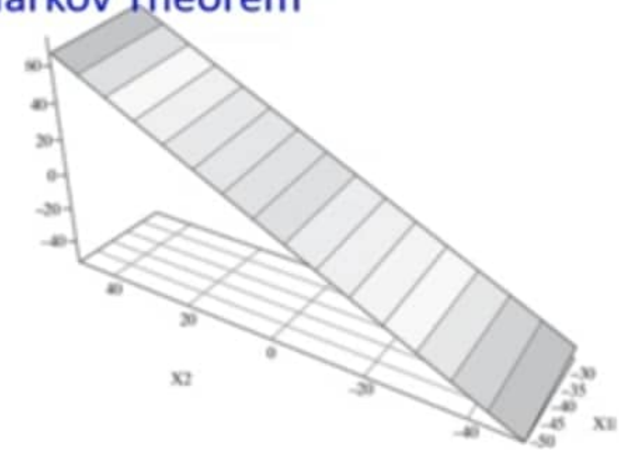


$\boxed{b = 1.89}$

- Slope measures the estimated change in the average value of Y as a result of a one-unit change in X.
- Here, slope = 1.89 tells us that the average value of the external examination marks increases by $\boxed{1.89}$ for each additional 1 mark in the internal examination.

Machine Learning
Subject Code: 20A05602T
UNIT IV - Supervised Learning: Regression – 4-1-3
Multiple Linear Regression

- Partial Regression Coefficients
- Best Linear Unbiased Estimator (BLUE) or Gauss-Markov Theorem
- Problems in Multi Linear Regression Analysis
  - Multicollinearity
  - Heteroskedasticity
- Bias and Variance in Regression Model
- Improving Accuracy
  - 1. Shrinkage Approach
  - 2. Subset Selection
  - 3. Dimensionality (Variable) Reduction



Multiple Linear Regression

- Regression is finding a relationship between the dependent variable (Y) and the independent variable(s) (X), $Y = f(X) = a + bX$ slope y intercept
- **In a multiple regression model, the target variable (dependent variable Y) is dependent on two or more independent variables (X1,X2,…,Xn)**
- **The dependent variable (Y) is continuous.**
- The equation for the relationship with two predictor variables, namely X1 and X2.
- $\hat{Y} = a + b1\ X1 + b2\ X2$

Subscribe

Multiple Linear Regression- Intercepts & Partial regression coefficients

- $\hat{Y} = a + b1\ X1 + b2\ X2$
- The model describes a plane in the three-dimensional space of $\hat{Y}$, X1, and X2.
- Parameter 'a' is the intercept of this plane.
- Parameters 'b1' and 'b2' are referred to as partial regression coefficients.
- Parameter b1 represents the change in the mean response corresponding to a unit change in X1 when X2 is held constant.
- Parameter b2 represents the change in the mean response corresponding to a unit change in X2 when X1 is held constant.

Example

home

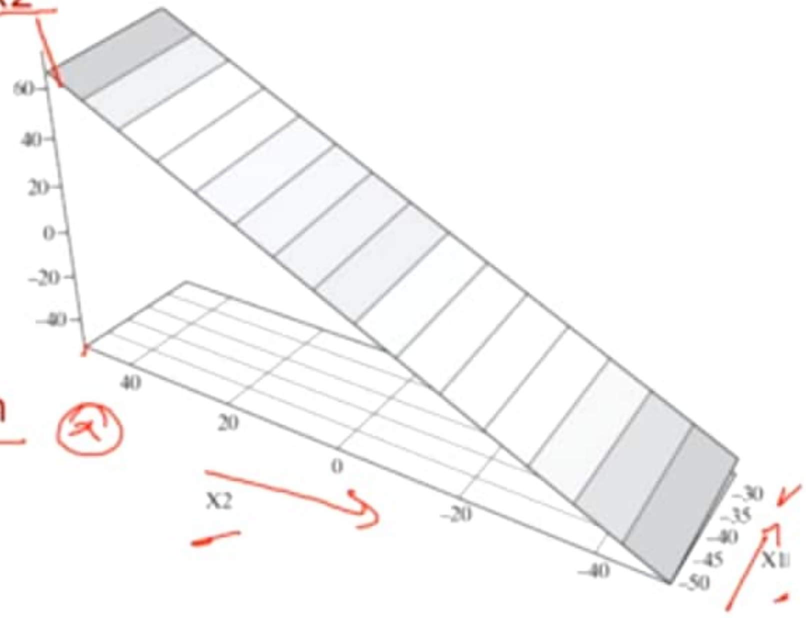
- **Karen's real estate problem,**
- Dependent variable is Price of a Property

$y =$ rate

- Independent variables (Predictor variables) are Area of the Property (in sq. m.) location, floor, number of years since purchase, and etc.
- A multiple regression equation as shown below:
- $\text{Price}_{\text{Property}} = f(\text{Area}_{\text{Property}}, \text{location, floor, Ageing, Amenities})$

Multiple Linear Regression...

- A multiple linear regression model with two predictor variables, namely X1 and X2
- $\boxed{\hat{Y} = a + b1X1 + b2X2}$
- $\hat{Y} = 22 + 0.3X1 + 1.2X2$
- Multiple regression for estimating equation when there are 'n' predictor variables is as follows:
- $\hat{Y} = a + b1\ X1 + b2\ X2 + b3\ X4 + \dots + bn\ Xn$
- While finding the best fit line, we can fit either a polynomial or curvilinear regression.



Best Linear Unbiased Estimator (BLUE) or Gauss-Markov Theorem or Assumptions in Regression Analysis

8

1. The dependent variable (Y) can be calculated / predicated as a linear function of a specific set of independent variables (X's) plus an error term (ε).

   $Y = f(x) + e$

2. The number of observations (n) is greater than the number of parameters (k) to be estimated, i.e. **n>k**.
3. Relationships determined by regression are only relationships of association based on the data set.
4. Regression line can be valid only over a limited range of data. If the line is extended (outside the range of extrapolation), it may only lead to wrong predictions.

Best Linear Unbiased Estimator (BLUE) or Gauss-Markov Theorem or Assumptions in Regression Analysis...

5. If the business conditions change and the business assumptions underlying the regression model are no longer valid, then the past data set will no longer be able to predict future trends.
6. Variance is the same for all values of X (homoskedasticity).
7. The error term (ε) is normally distributed. This also means that the mean of the error (ε) has an expected value of 0. ← no error.
8. The values of the error (ε) are independent and are not related to any values of X. This means that there are no relationships between a particular X, Y that are related to another specific value of X, Y.

Main Problems in Multi Linear Regression Analysis

- Two primary problems:

$x_1, x_2 \cdots x_n$

$x_3, x_n \cdots$

- 1. Multicollinearity
- Several independent variables in a model are corelated.
- Two are more independent variables are strongly corelated.
- 2. Heteroskedasticity.
- Unequal scattering of data points i.e. The observed values deviate from the predicted values ununiformly.

Bias and Variance in Regression Model

- The bias and variance is similar to accuracy and prediction.
- Accuracy refers to how close the estimation is near the actual value,
- Prediction refers to continuous estimation of the value.

- High bias = low accuracy (not close to real value)

- High variance = low prediction (values are scattered)
- Low bias = high accuracy (close to real value)

- Low variance = high prediction (values are close to each other)

Bias and Variance in Regression Model...

- A regression model which is highly accurate and highly predictive; then
- the overall error of our model will be low,
- a low bias (high accuracy) and low variance (high prediction).
- This is highly preferable.
- If the **variance increases** (low prediction), the spread of our data points increases, which results in less accurate prediction.
- As the **bias increases** (low accuracy), the error between our predicted value and the observed values increases.
- Therefore, balancing out bias and accuracy is essential in a regression model.

Improving Accuracy of Linear Regression

- Accuracy of linear regression can be improved using the following three methods:
- 1. Shrinkage Approach
- 2. Subset Selection
- 3. Dimensionality (Variable) Reduction

Shrinkage (Regularization) approach

- By limiting (shrinking) the estimated coefficients, we can try to reduce the variance at the cost of a **negligible increase in bias.**
- This can improve the accuracy of the model.
- Few variables used in the multiple regression model are not associated with the overall response, are called as irrelevant variables;
- this may lead to unnecessary complexity in the regression model.

Shrinkage (Regularization) approach...

- The two best-known techniques for shrinking the regression coefficients towards zero are
  - 1. Ridge Regression
  - 2. lasso (Least Absolute Shrinkage Selector Operator)

Subscribe

Subset Selection

- **Identify a subset of the predictors** that is assumed to be related to the response.
- Fit a model using OLS on the selected reduced subset of variables.
- There are two methods in which subset of the regression can be selected:
- 1. Best subset selection
- 2. Stepwise subset selection
  - 1. Forward stepwise selection (0 to k)
  - 2. Backward stepwise selection (k to 0)

1. Best subset selection

- In this, fit a separate least squares regression for each possible subset of the k predictors.
- The best subset selection procedure considers all the possible (2^k) models containing subsets of the p predictors.

2. Stepwise subset selection

- The stepwise subset selection method can be applied to choose the best subset.
- There are two stepwise subset selection:
  - 1. Forward stepwise selection (0 to k)
  - 2. Backward stepwise selection (k to 0)

Subset selection – Forward stepwise selection

- Forward stepwise selection is a computationally efficient alternative to best subset selection.
- It begins with a model containing no predictors, and then, predictors are added one by one to the model, until all the k predictors are included in the model.
- At each step, the variable (X) that gives the highest additional improvement to the fit is added.

Subset selection – Backward stepwise selection

- Backward stepwise selection begins with the least squares model which contains all k predictors and then iteratively removes the least useful predictor one by one.

3 Dimensionality Reduction (Variable Reduction)

- The number of variables is reduced using the dimensionality reduction method.
- In dimensionality reduction, predictors (X) are transformed, and the model is set up using the transformed variables after dimensionality reduction.
- Principal component analysis is one of the most important dimensionality (variable) reduction techniques.

PCA

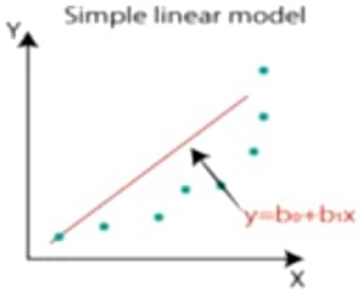
Machine Learning
Subject Code: 20A05602T

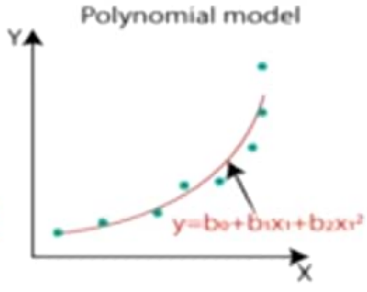
UNIT IV - Supervised Learning: Regression – 4-1-4

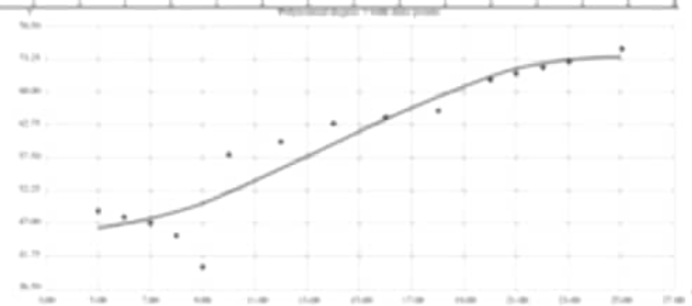
Polynomial Regression Model

- Example - Predicting External Marks based on Internal Marks for 15 students

Internal Exam (X)	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam (Y)	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68







Polynomial Regression Model $y = f(x)$

- Polynomial regression model is the extension of the simple linear model by adding extra predictors obtained by raising the power of predictors
- For example, if there are three variables, then X, X^2, and X^3 are used as predictors.
- This approach provides a simple way to yield a non-linear fit to data.

$$y = f(x) = c_0 + c_1.X^1 + c_2.X^2 + c_3.X^3$$

- Where c_0, c_1, c_2, and c_3 are the coefficients.
- It is also called the special case of Multiple Linear Regression, Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

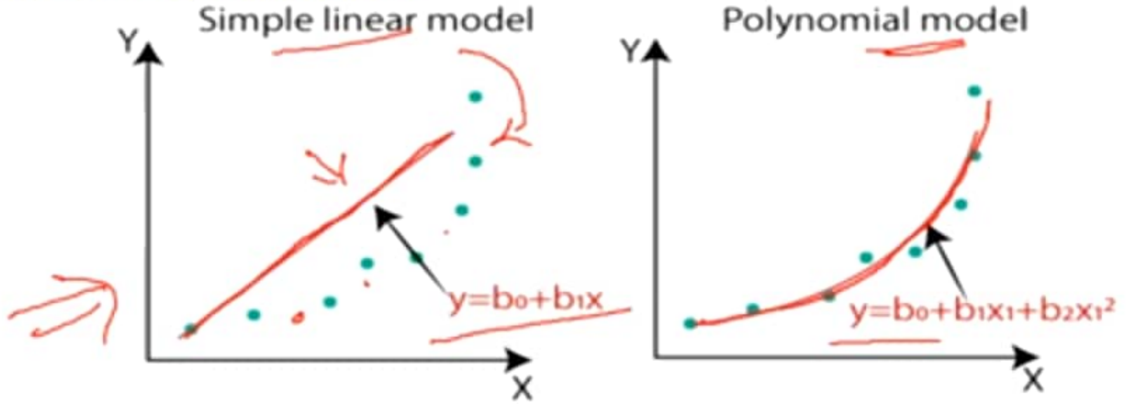
Polynomial Regression Model...

- Polynomial Regression is a regression algorithm that models the relationship between a dependent(Y) and independent variable(X) as nth degree polynomial.
- The Polynomial Regression equation is given below
- $Y = c_0 + c_1X^1 + c_2X^2 + \ldots + c_nX^n$
- It is a linear model with some modification in order to increase the accuracy.
- The non linear nature of dataset used in Polynomial regression for training the model.
- It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.

Polynomial Regression Model...

$y = a + bx$

- If we apply a linear model on a linear dataset, then it provides us a good result as we have seen in Simple Linear Regression.
- if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output.
- Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.
- Hence if the data points are arranged in a non-linear fashion, we need the Polynomial Regression model.



Equation of the Polynomial Regression Model:

$$y = b_0 x_0^0 + b_1 x_1^1$$

- Simple Linear Regression : $y = b_0+b_1x^1$
- Multiple Linear Regression: $y = b_0+b_1x+ b_2x_2+ b_3x_3+....+ b_nx_n$
- Polynomial Regression : $y = b_0+b_1x + b_2x^2+ b_3x^3+....+ b_nx^n$

Subscribe

Example – Predicting External Marks based on Internal Marks for 15 students

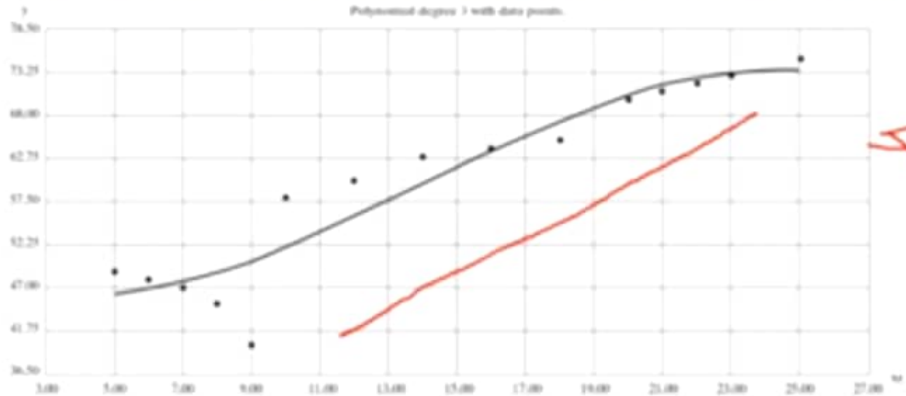
- Let us use the below data set of (X, Y) for degree 3 polynomial

TDS

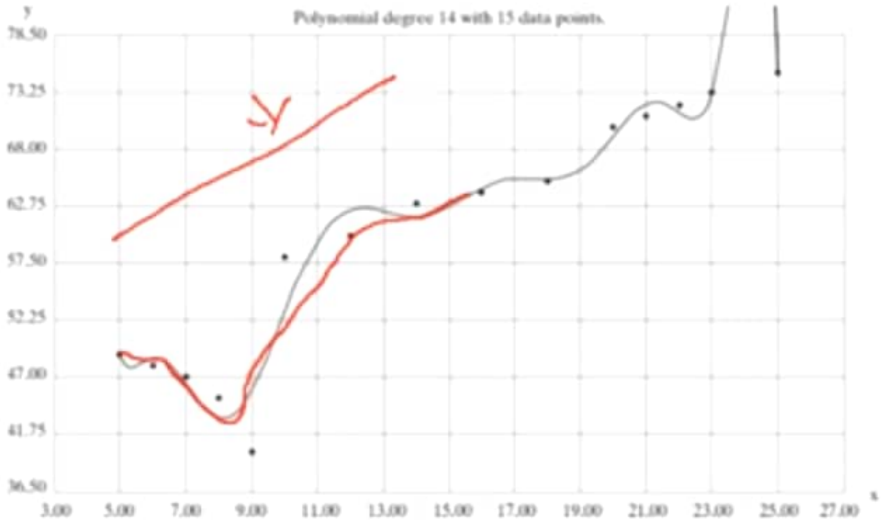
Internal Exam (X)	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam (Y)	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

3

- the regression line is slightly curved for polynomial degree 3 with the above 15 data points. (Polynomial regression degree 3)



- The regression line will curve further if we increase the polynomial degree.
- At the extreme value as shown below, the regression line will be **overfitting** into all the original values of X. Polynomial regression degree 14



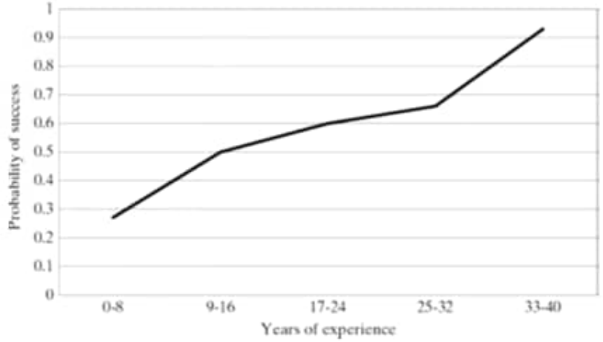
Machine Learning
Subject Code: 20A05602T

UNIT IV - Supervised Learning: Regression – 4-1-5

Logistic Regression Model

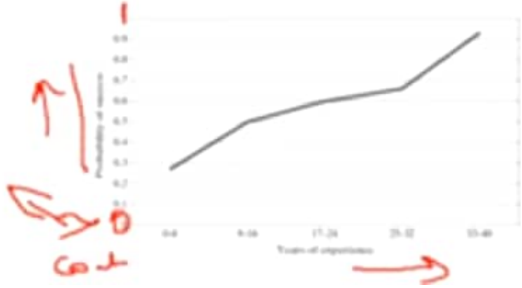
- Example –
- Assumptions in logistic regression

X	Y
0–8	0.27
9–16	0.5
17–24	0.6
25–32	0.66
33–40	0.93



Logistic Regression

- In logistic regression, dependent variable (Y) is binary (0,1) and independent variables (X) are continuous in nature.
- If X and Y have a strong positive linear relationship, the probability of Y = 1 will increase as values of X increase.
- The goal of logistic regression is to predict the likelihood that Y is equal to 1 (**probability that Y = 1, rather than 0**) given certain values of X.
- So, we are **predicting probabilities rather than the scores** of the dependent variable.
- Logistic regression is used in both classification and regression technique depending on the scenario.
- It is used for predicting the outcome of a categorical dependent variable similar to OLS regression.



Logistic Regression - Example

- Predict a small project will succeed or fail on the basis of the number of years of experience of the project manager handling the project.
- The project managers who have been managing projects for many years will be more likely to succeed..
- This means that as X (the number of years of experience of project manager) increases, the probability that Y will be equal to 1 (success of the new project) will tend to increase.

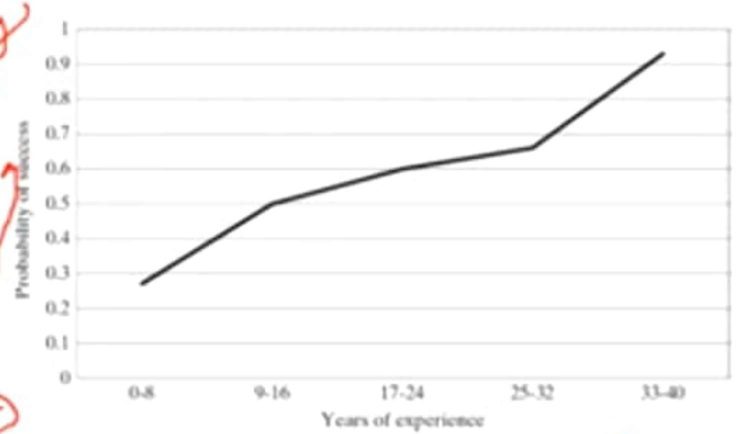
Logistic Regression – Example... Training data Set

- If 60 already executed projects were studied, and the years of experience of project managers ranges from 0 to 20 years, then increase the probability that Y = 1 with a graph.
- Years of experience into categories (i.e. 0–8, 9–16, 17–24, 25–32, 33–40).
- If we compute the mean score on Y (averaging the 0s and 1s) for each category of years of experience,

T DS

X	Y
0–8	0.27
9–16	0.5
17–24	0.6
25–32	0.66
33–40	0.93

y

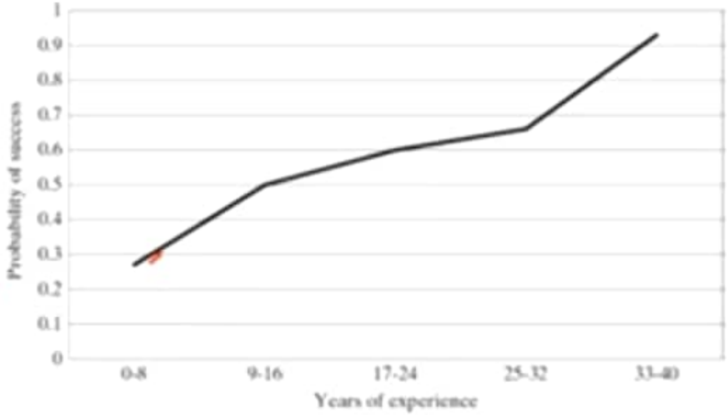


x

Logistic Regression - Example...

- As X increases, the probability that $Y = 1$ increases.
- If the project manager has more years of experience, success rate of projects increased.
- A perfect relationship represents a perfectly curved S rather than a straight line, as was the case in OLS regression.
- So, to model this relationship, we need some fancy algebra / mathematics that accounts for the bends in the curve.

X	Y
0–8	0.27
9–16	0.5
17–24	0.6
25–32	0.66
33–40	0.93



Logistic Regression...

S

log

- The logistic regression begins with an explanation of the logistic function, which always takes values between zero and one.

0 - 1

- The logistic formulae are stated in terms of the probability that Y = 1, which is referred to as P.

P : Y=1

- The probability that Y is 0 is 1 – P.

Y=0 - 1 - P

$$\ln\left(\frac{P}{1-P}\right) = \boxed{a + bX} \leftarrow Y = 1$$

$$\text{log} \quad \ln(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \boxed{\varepsilon}$$

- The 'ln' symbol refers to a natural logarithm and a + bX is the regression line equation.
- Probability (P) can also be computed from the regression equation.

Logistic Regression...

- calculate the expected probability that $\boxed{Y = 1}$ for a given value of X.

$$P = \frac{\exp(a + bX)}{1 + \exp(a + bx)} = \boxed{\frac{e^{a+bx}}{1 + e^{a+bx}}}$$

- 'exp' is the exponent function, which is sometimes also written as e.

Logistic Regression – Example-2

- Let us say we have a model that can predict whether a person is male or female on the basis of their height.
- Given a height of 150 cm, we need to predict whether the person is male or female.
- We know that the coefficients of a = −100 and b = 0.6.
- Using the above equation, we can calculate the probability of male given a height of 150 cm or more formally P(male|height = 150).
- y = e^(a + b × X)/(1 + e^(a + b × X))
- y = exp (−100 + 0.6 × 150)/(1 + EXP(−100 + 0.6 × X)
- y = 0.000046 ≈ 0
- or a probability of near zero that the person is a **male**. (the person is female)

male

female

Assumptions in logistic regression

- There exists a linear relationship between logic function and independent variables
- The dependent variable Y must be categorical (1/0) and take binary value,
- e.g. if pass then Y = 1; else Y = 0
- The data meets the independent and identically distributed (iid) criterion,
- i.e. the error terms ε, are independent from one another and identically distributed.
- The error term follows a binomial distribution [n, p]
- where
  - n = # of records in the data
  - p = probability of success (pass, responder)

Subscribe

Machine Learning
Subject Code: 20A05602T

UNIT IV - Supervised Learning: Regression – 4-1-6

Maximum Likelihood Estimation (MLE)

- Definition and Example

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Maximum Likelihood Estimation

- Maximum Likelihood Estimation (MLE) used to estimate the coefficient in logistic regression
- A coin flips problem, outcome equally heads and tails of the same number of times.
- If we toss the coin 10 times, it is expected that we get five times Head and five times Tail.
- Let us now discuss about the probability of getting only Head as an outcome;
- it is 5/10 = 0.5 in the above case.
- If P is greater than 0.5, it is in favour of Head,
- P is lesser than 0.5, it is against the Head.

Maximum Likelihood Estimation...

- Let us represent 'n' flips of coin as X1 , X2 , X 3,..., Xn .
- Now Xi can take the value of 1 or 0. t h
- Xi = 1 if Head is the outcome
- Xi = 0 if Tail is the outcome
- When we use the Bernoulli distribution represents each flip of the coin:

$$\boxed{f(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1 - x_i}}$$

Maximum Likelihood Estimation...

- Each observation $\boxed{X}$ is independent and identically distributed (iid), and the joint distribution simplifies to a product of distributions.

$$f(x_1, \ldots, x_n|\theta) = {}_{i=1}^{n} f(x_i|\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \ldots \theta^{x_n}(1-\theta)^{1-x_n} = \boxed{\theta^{\#H}(1-\theta)^{n-\#H}},$$

- where #H is the number of flips that resulted in the expected outcome (heads in this case).
- The likelihood equation is

$$\boxed{L(\theta|x) = {}_{i=1}^{n} f(x_i|\theta)}$$

Maximum Likelihood Estimation...

- But the likelihood function is not a probability.
- The likelihood for some coins may be 0.25 or 0 or 1.
- MLE is about predicting the value for the parameters that maximizes the likelihood function.

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$