UNIT IV Supervised Learning: Regression Lecture 10Hrs
Introduction, Example of Regression, Common Regression Algorithms-Simple linear regression, Multiple linear regression, Assumptions in Regression Analysis, Main Problems in Regression Analysis, Improving Accuracy of the Linear Regression Model, Polynomial Regression Model, Logistic Regression, Maximum Likelihood Estimation.

Supervised Learning: Regression

8.1INTRODUCTION

- **Regression in machine learning consists of mathematical methods that allow data scientists to predict a continuous outcome (y) based on the value of one or more predictor variables (x). Linear regression is probably the most popular form of regression analysis because of its ease-of-use in predicting and forecasting.**
- A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points

8.2 EXAMPLE OF REGRESSION

- In the context of regression, dependent variable (Y) is the one whose value is to be predicted, e.g. the price quote of the real estate. This variable is presumed to be functionally related to one (say, X) or more independent variables called predictors.
- In real estate, area of the property, location, floor, etc. as predictors of the model can be used. In other words, the dependent variable depends on independent variable(s) or predictor(s).
- Linear Regression is applicable in real-world scenarios where machine learning problems can be used to predict the output as a continuous variable.
- In agriculture, Linear Regression can be used to predict the amount of rainfall and crop yield, while in banking, it is implemented to predict the probability of loan defaults. For the Finance sector, Linear Regression is used to predict stock prices and assess associated risks. In the healthcare sector, Linear Regression is helpful in modeling healthcare costs, predicting the length of stay in hospitals for patients, etc. In the domain of sports analytics, Linear Regression can be used to predict the performance of players in upcoming games. Similarly, it can be used in education to predict student performances in different courses. Businesses also use Linear Regression to forecast product demands, predict product sales, decide on marketing and advertising strategies, and so on.

8.3COMMON REGRESSION ALGORITHMS

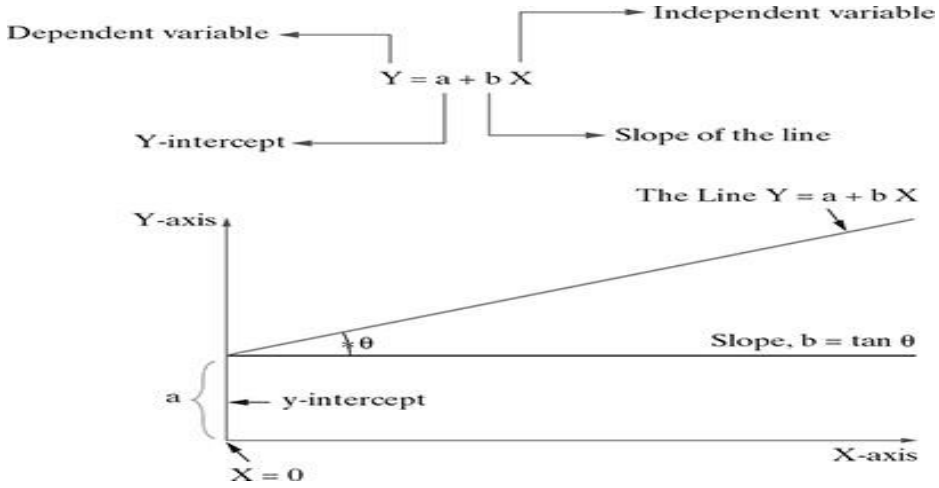
The most common regression algorithms are

- Simple linear regression Multiple linear regression Polynomial regression
- Multivariate adaptive regression splines Logistic regression
- Maximum likelihood estimation (least squares)

8.3.1 Simple Linear Regression

- As the name indicates, simple linear regression is the simplest regression model which involves only one predictor. This model assumes a linear relationship between the dependent variable and the predictorvariable

Simple linear regression



- In the context of Karen's problem, if we take Price of a Property as the dependent variable and the Area of the Property (in sq. m.) as the predictor variable, we can build a model using simple linear regression.

$$\text{Price}_{\text{Property}} = f(\text{Area}_{\text{Property}})$$

- Assuming a linear association, we can reformulate themodel as

$$\text{Price}_{\text{Property}} = a + b.\ \text{Area}_{\text{Property}}$$

where '*a*' and '*b*' are intercept and slope of the straightline, respectively.

- Just to recall, straight lines can be defined in a slope– intercept form $Y = (a + bX)$, where a = intercept and b = slope of the straight line. The value of intercept indicates the value of Y when $X = 0$. It is known as 'the intercept or Y intercept' because it specifies where the straight line crosses the vertical or Y-axis (refer to above figure).

8.3.1.1 Slope of the simple linear regression model

- Slope of a straight line represents how much the line in a graph changes in the vertical direction (Y-axis) over a change in the horizontal direction (X-axis) as shown in Figure 8.2.

Slope = Change in Y/Change in X

- Rise is the change in Y-axis ($Y_2 - Y_1$) and Run is the change in X-axis ($X_2 - X_1$). So, slope is represented asgiven below:

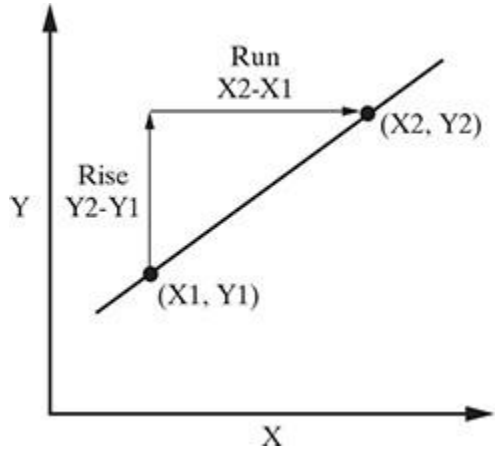


FIG. 8.2 Rise and run representation

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

Example of slope

- Let us find the slope of the graph where the lower pointon the line is represented as (−3, −2) and the higher point on the line is represented as (2, 2).

  $(X_1, Y_1) = (-3, -2)$ and $(X_2, Y_2) = (2, 2)$
  Rise = $(Y_2 - Y_1) = (2 - (-2)) = 2 + 2 = 4$
  Run = $(X_2 - X_1) = (2 - (-3)) = 2 + 3 = 5$
  Slope = Rise/Run = 4/5 = 0.8
- There can be two types of slopes in a linear regression model: positive slope and negative slope. Different types of regression lines based on the **type of slope include(5 Marks)**
  - Linear positive slope
  - Curve linear positive slope
  - Linear negative slope
  - Curve linear negative slope

Linear positive slope:

A positive slope always moves upward on a graph fromleft to right (refer to Fig. 8.3).

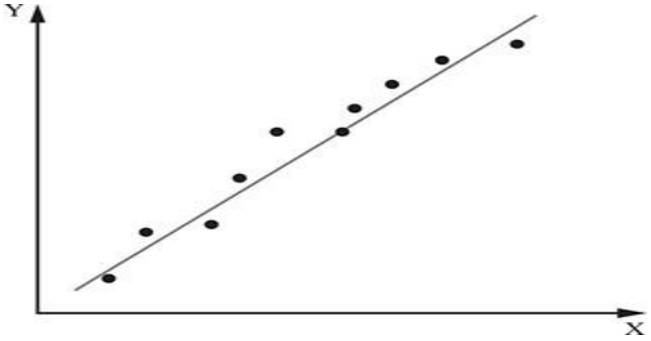


FIG. 8.3 Linear positive slope

Slope = Rise/Run = $(Y2 - Y1) / (X2 - X1)$ = Delta (Y) / Delta(X)

- Scenario 1 for positive slope: Delta (Y) is positive and Delta (X) ispositive
- Scenario 2 for positive slope: Delta (Y) is negative and Delta (X) isnegative

Curve linear positive slope

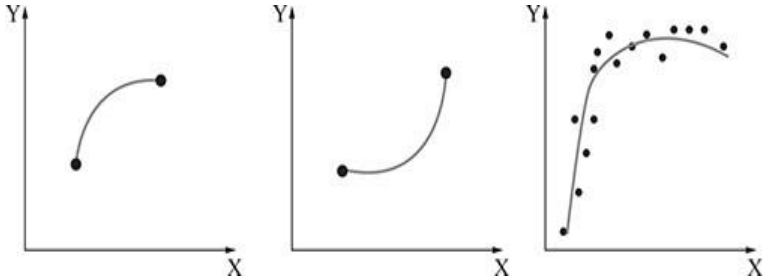


FIG. 8.4 Curve linear positive slope

- Curves in these graphs (refer to Fig. 8.4) slope upwardfrom left to right.

  Slope = $(Y_2 - Y_1) / (X_2 - X_1)$ = Delta (Y) / Delta(X)
- Slope for a variable (X) may vary between two graphs, but it will always be positive; hence, the above graphs arecalled 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 (refer to Fig. 8.5).

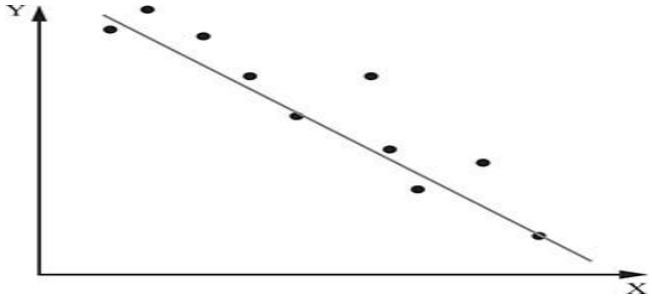


FIG. 8.5 Linear negative slope

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

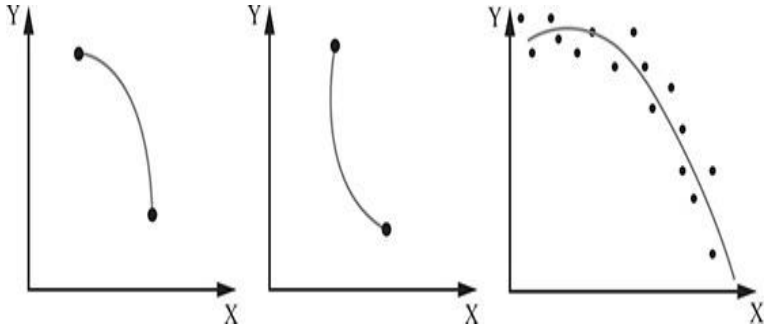


FIG. 8.6 Curve linear negative slope

- Curves in these graphs (refer to Fig. 8.6) slope downward from left to right.

  Slope = $(Y_2 - Y_1) / (X_2 - X_1)$ = 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.

8.3.1.2 No relationship graph

- Scatter graph shown in Figure 8.7 indicates 'no relationship' curve as it is very difficult to conclude whether the relationship between *X* and *Y* is positive or negative.

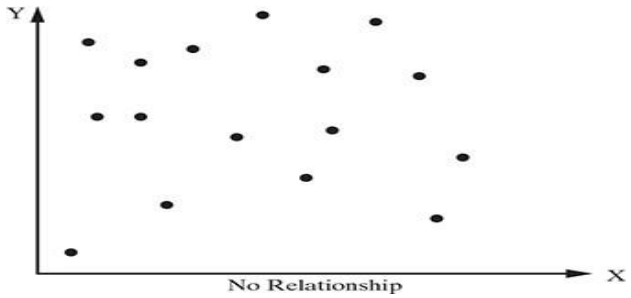


FIG. 8.7 No relationship graph

8.3.1.3 Error in simple regression:

- The regression equation model in machine learning uses the above slope–intercept format in algorithms. 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. However, 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.**

$$Y = (a + bX) + \varepsilon.$$

8.3.1.4 Example of simple regression

- A college professor believes that if the grade for internalexamination 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 givenbelow:

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 as depicted in Figure 8.8.

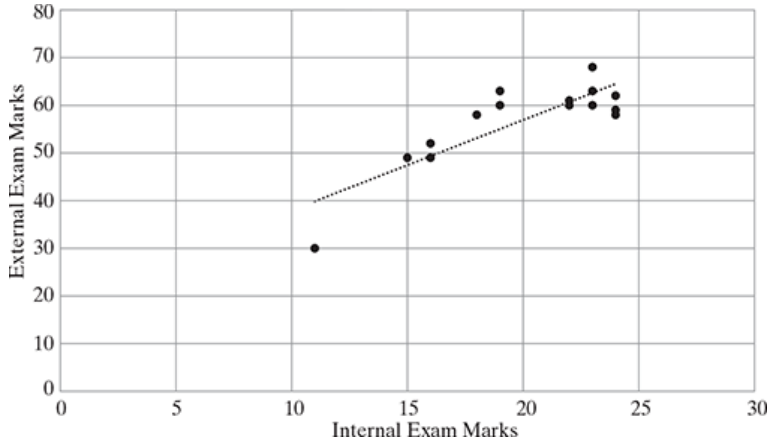


FIG. 8.8 Scatter plot and regression line

- As you can observe from the above graph, the line (i.e.the regression line) does not predict the data exactly (refer to Fig. 8.8). 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 as depicted in Figure 8.9.

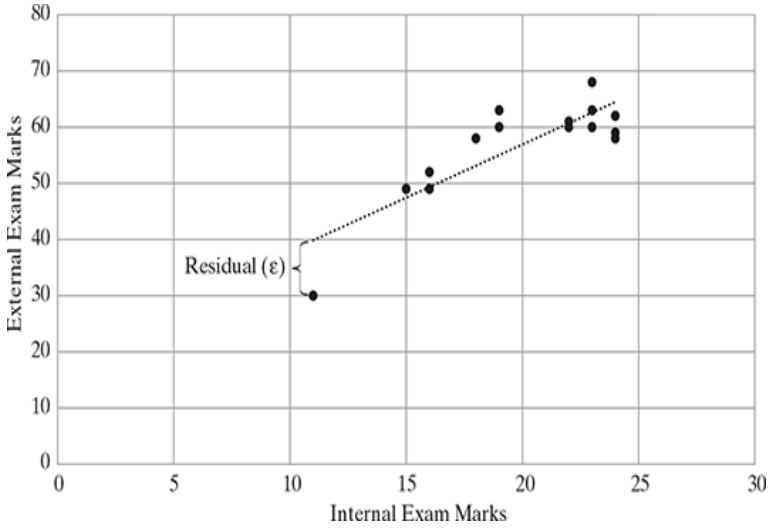


FIG. 8.9 Residual error

- As we know, in simple linear regression, the line isdrawn using the regression formula.

$$Y = (a + bX) + \varepsilon$$

- If we know the values of '*a*' and '*b*', then it is easy to predict the value of *Y* for any given *X* by using the above formula. But the question is how to calculate the values of '*a*' and '*b*' for a given set of *X* and *Y* values?
- A straight line is drawn as close as possible over the points on the scatter plot. 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 means summing the errors of each prediction or, more appropriately, the Sum of the Squares of the Errors

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

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

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

- The corresponding value of '*a*' calculated using theabove value of '*b*' is

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

- So, let us calculate the value of *a* and *b* for the given example. For detailed calculation, refer to Figure 8.10.

Calculation summary

Sum of X = 299Sum of Y = 852

Mean X, M_X = 19.93Mean Y, M_Y = 56.8

Sum of squares (SS_X) = 226.9333Sum of products (SP) = 429.8 Regression equation = $\hat{y} = bX + a$

$$b = \frac{\text{SP}}{\text{SS}_X} = \frac{429.8}{226.93} = 1.89395$$

$a = M_Y - bM_X = 56.8 - (1.89 \times 19.93) = 19.0473$

$\hat{y} = 1.89395X + 19.0473$

Hence, for the above example, the estimated regression equation is constructed on the basis of the estimated values of *a* and *b*:

$$\hat{y} = 1.89395X + 19.0473$$

So, in the context of the given problem, we can say

Marks in external exam = 19.04 + 1.89 × (Marks in internal exam)

or, $M_{Ext} = 19.04 + 1.89 \times M_{Int}$

FIG. 8.10 Detailed calculation of regression parameters

Step 2 Step 3 Step 5

X	Y	X- mean (X)	Y- Mean (Y)	$(Xi - \overline{X})(Yi - \overline{Y})$	$(Xi - \overline{X})^2$
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	34.384	9.4249
19.9	**56.8**		$\Sigma(Xi - \overline{X})(Yi - \overline{Y})$	**429.8**	**226.9335**

Step 1 Step 4 Step 6

Step 7: Divide (step4 / step6)

b = 429.28 / 226.93 = 1.89

Step 8: Calculate a using the value of b

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

a = 56.8 – 1.89 × 19.9

a = 19.05

The model built above can be represented graphically as

- an extended version (refer to Fig. 8.11) a zoom-in version (refer to Fig. 8.12)

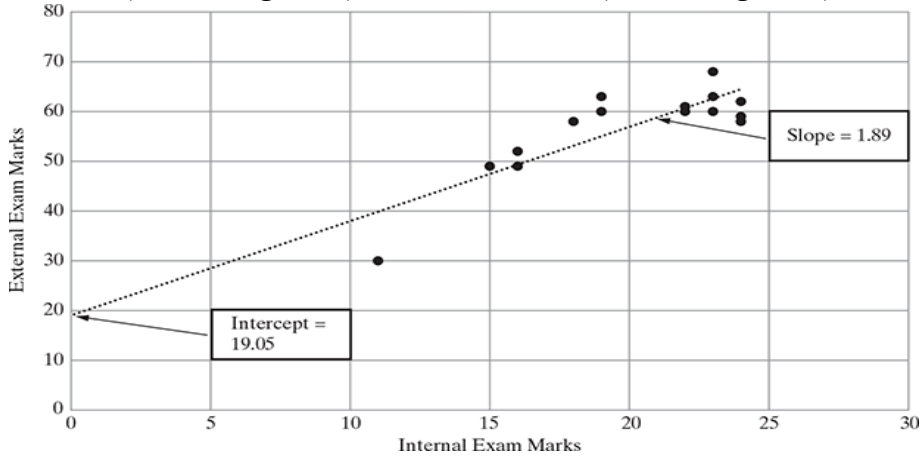


FIG. 8.11 Extended version of the regression graph

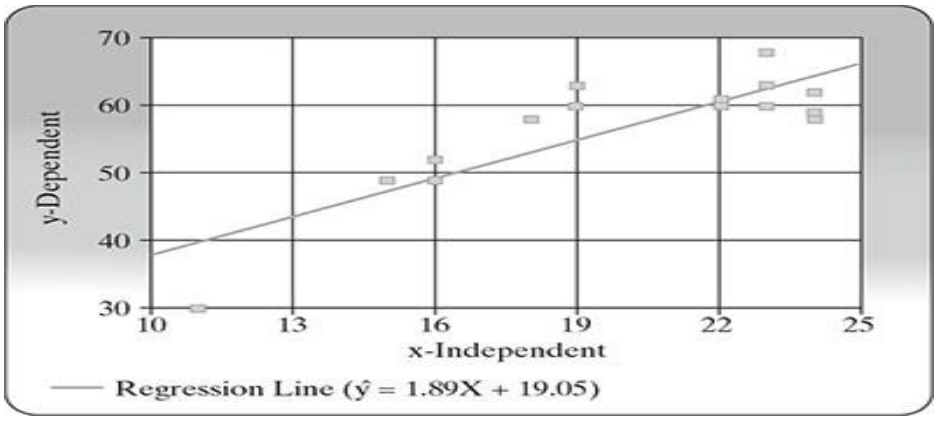


FIG. 8.12 Zoom-in regression line

Interpretation of the intercept:

- The simple linear regression model built on the data in the example is

$$M_{\mathbf{Ext}} = \mathbf{19.04} + \mathbf{1.89} \times M_{\mathbf{Int}}$$

- The value of the intercept from the above equation is 19.05. However, none of the internal mark is 0. So, intercept = 19.05 indicates that 19.05 is the portion of the external examination marks not explained by the internal examination marks.
- 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 1.89 for each additional 1 mark in the internal examination.
- Now that we have a complete understanding of how to build a simple linear regression model for a given problem, it is time to summarize the algorithm.

8.3.1.5 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'

8.3.1.6 Maximum and minimum point of curves

- Maximum (shown in Fig. 8.13) and minimum points (shown in Fig. 8.14) on a graph are found at points where the slope of the curve is zero. It becomes zero either from positive or negative value. The maximum point is the point on the curve of the graph with the highest *y*-coordinate and a slope of zero. The minimum point is the point on the curve of the graph with the lowest *y*- coordinate and a slope of zero.

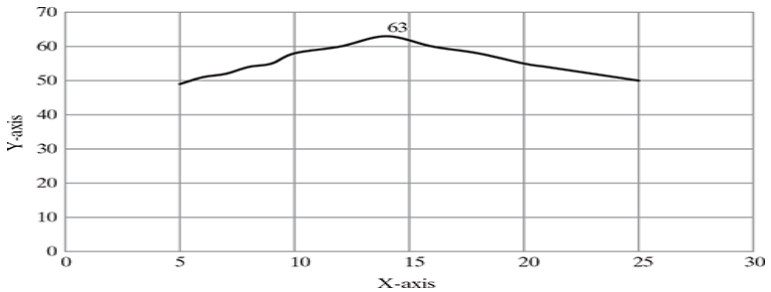


FIG. 8.13 Maximum point of curve

- Point 63 is at the maximum point for this curve (refer to Fig. 8.13). Point 63 is at the highest point on this curve. It has a greater *y*-coordinate value than any other point on the curve and

has a slope of zero.

- Point 40 (marked with an arrow in Fig. 8.14) is the minimum point for this curve. Point 40 is at the lowest point on this curve. It has a lesser *y*-coordinate value than any other point on the curve and has a slope of zero.

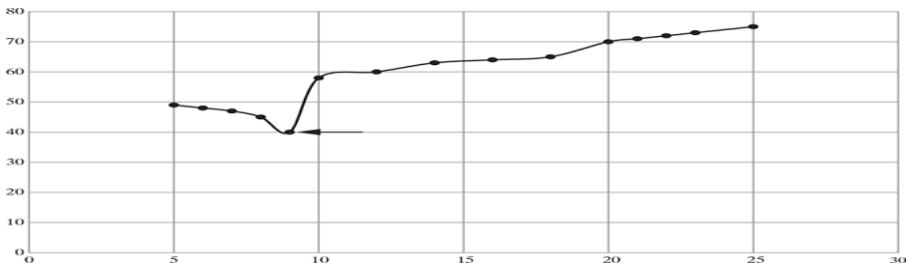


FIG. 8.14 Minimum point of curve

8.3.2 Multiple Linear Regression

- Moreover, Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. **We can define it as:**
- **Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.**

Multiple linear regression formula

- The simple linear regression model and the multiple regression model assume that the dependent variable iscontinuous.
- The following expression describes the equation involving the relationship with two predictor variables,namely X_1 and X_2.

$$Y = a + b1X1 + b2X2$$

- The model describes a plane in the three-dimensionalspace of Y, X_1, and X_2. Parameter 'a' is the intercept of this plane. Parameters 'b_1' and 'b_2' are referred to as **partial regression coefficients**..

- Multiple regression for estimating equation when there are 'n' predictor variables is as follows:

  $y= a+b_1 X_1+b_2 X_2+\ldots\ldots+b_n Xn$

  y= the predicted value of the dependent variable

  a= the y-intercept

  $b_1 X_1$= the regression coefficient (1) of the first independent variable (X1)

  bn Xn= the regression coefficient of the last independent variable

- Consider the following example of a multiple linear regression model with two predictor variables, namely $X_1$and X_2 (refer to Fig. 8.15).

$$Y = 22 + 0.3X_1 + 1.2X_2$$

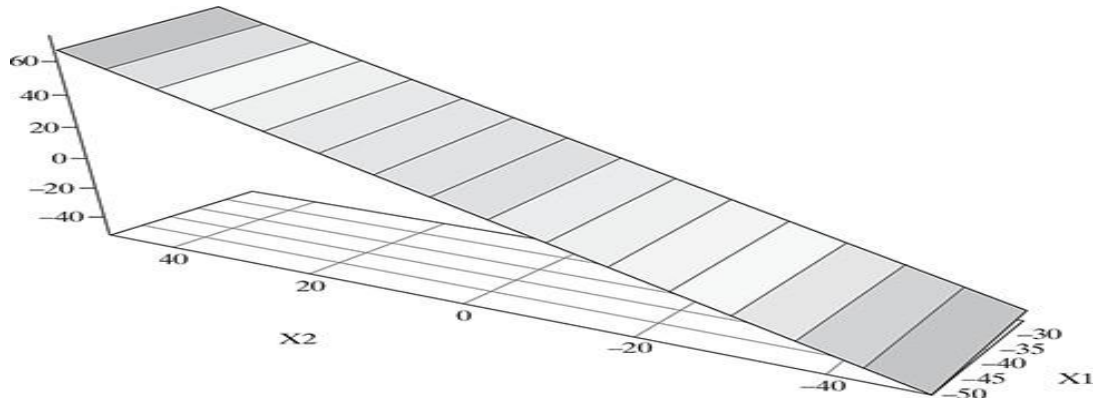


FIG. 8.15 Multiple regression plane

8.3.3 Assumptions in Regression Analysis

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).
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 and not necessarily of cause and effect of the defined class.
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.
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.
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*.

Given the above assumptions, the OLS estimator is the **Best Linear Unbiased Estimator (BLUE)**, and thisis called as **Gauss-Markov Theorem**.

8.3.4 Main Problems in Regression Analysis

In multiple regressions, there are two primary problems:multicollinearity and heteroskedasticity.

8.3.4.1 Multicollinearity

- Two variables are perfectly collinear if there is an exact linear relationship between them. **Multicollinearity is thesituation** in which the degree of correlation is not only between the dependent variable and the independent variable, but there is also a strong correlation within (among) the independent variables themselves.
- A multiple regression equation can make good predictions when there is multicollinearity, but it is difficult for us to determine how the dependent variable will change if each independent variable is changed one at a time.

Effects of Multicollinearity

- Multicollinearity has unpleasant effects which are highlighted below:
  - The independent variables may assume each other's role
  - The effect of the independent variables on the dependent variable cannot be distinguished
  - Estimation of the regression coefficients is rendered unreliable
  - In some cases, analysis becomes very difficult to perform

Measure of Multicollinearity

- There are three metrics used to examine the effects of multicollinearity. They are:
  - VIF (Variation Inflation Factor)
  - Tolerance
  - Condition Indices
- **Variation Inflation Factor (VIF):** Variance Inflation Factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if the predictors are correlated. If no factors arecorrelated, the VIFs will be equal to 1.
  - VIF is given by the formula

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

- **Tolerance:** This is given as the inverse of the Variation Inflation factor. A low tolerance value indicates an increasing multicollinearity

$$T_i = 1 - {R_i}^2$$

- **Condition Index:** This is a measure of the relative amount of variation associated with an eigenvalue. Large value of the condition index indicates a high degree of collinearity.

8.3.4.2 Heteroskedasticity

- Heteroskedasticity refers to the changing variance of the error term. If the variance of the error term is not constant across data sets, there will be erroneous predictions. In general, for a regression equation to make accurate predictions, the error term should be independent, identically (normally) distributed (iid).

- Mathematically, this assumption is written as

$$\text{var}(u_i|X) = \sigma^2 \quad \text{and}$$
$$\text{cov}(u_i u_j|X) = 0 \quad \text{for } i \neq j$$

  - where 'var' represents the variance, 'cov' represents the covariance, '*u*' represents the error terms, and '*X*' represents the independent variables.
- This assumption is more commonly written as

$$\text{var}(u_i) = \sigma^2 \quad \text{and}$$
$$\text{cov}(u_i u_j) = 0 \quad \text{for } i \neq j.$$

8.3.5 Improving Accuracy of the Linear Regression Model

- The concept of bias and variance is similar to accuracy and prediction. Accuracy refers to how close the estimation is near the actual value, whereas 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)**
- Let us say we have a regression model which is **highly accurate and highly predictive**; therefore, the overall error of our model will be low, implying a low bias (high accuracy) and low variance (high prediction). This is highly preferable.
- Similarly, we can say that 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.

In the linear regression model, it is assumed that the number of observations (*n*) is greater than the number of parameters (*k*) to be estimated, i.e. $n > k$, and in that case, the least squares estimates tend to have low variance and hence will perform well on test observations.

However, if observations (*n*) is not much larger than parameters (*k*), then there can be high variability in the least squares fit, resulting in overfitting and leading to poor predictions.

If $k > n$, then linear regression is not usable. This also indicates infinite variance, and so, the method cannot be used at all.

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

8.3.5.1 Shrinkage (Regularization) approach:

- By limiting (shrinking) the estimated coefficients, we cantry to reduce the variance at the cost of a negligible increase in bias. This can in turn lead to substantial improvements in the accuracy of the model.
- Few variables used in the multiple regression model are in fact not associated with the overall response and are called as irrelevant variables; this may lead to unnecessary complexity in the regression model.
- This approach involves fitting a model involving all predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing the overall variance. Some of the coefficients may also be estimated to be exactly zero, thereby indirectly performing variable selection.
- The twobest-known techniques for shrinking the regression coefficients towards zero are
    1. ridge regression
    2. lasso (Least Absolute Shrinkage Selector Operator)

Ridge regression:

- Ridge regression performs L2 regularization, i.e. it adds penalty equivalent to square of the magnitude of coefficients

Minimization objective of ridge = LS Obj + α × (sum ofsquare of coefficients)

Ridge regression (include all k predictors in the final model) is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. If $k > n$, then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance. Thus, ridge regression works best in situations where the least squares estimates have high variance. One disadvantage with ridge regression is that it will include all k predictors in the final model. This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which thenumber of variables k is quite large. Ridge regression willperform better when the response is a function of many predictors, all with coefficients of roughly equal size.

Lasso regression:

- Lasso regression performs L1 regularization, i.e. itadds penalty equivalent to the absolute value of the magnitude of coefficients.

Minimization objective of ridge = LS Obj + α ×(absolute value of the magnitude of coefficients)

- The *lasso* overcomes this disadvantage by forcing some of the coefficients to zero value. We can say that the lasso yields sparse models (involving only subset) that are simpler as well as more interpretable. The lasso can be expected to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or equal to zero.

8.3.5.2 Subset selection:

- Identify a subset of the predictors that is assumed to be related to the response and then 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:

k

1. Best subset selection (considers all the possible (2^k))
2. Stepwise subset selection
    1. Forward stepwise selection (0 to k)
    2. Backward stepwise selection (k to 0)

In best subset selection, we fit a separate least squaresregression for each possible subset of the *k* predictors.

For computational reasons, best subset selection cannot be applied with very large value of predictors (*k*). The

best subset selection procedure considers all the possible (2^k) models containing subsets of the *p* predictors.

The stepwise subset selection method can be applied to choose the best subset. There are two stepwise subsetselection:

1. Forward stepwise selection (0 to *k*)
2. Backward stepwise selection (*k* to 0)

Forward stepwise selection is a computationally efficient alternative to best subset selection. Forward stepwise considers a much smaller set of models, that toostep by step, compared to best set selection. Forward stepwise selection 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. In particular, at each step, the variable (*X*) that gives the highest additional improvement to the fit is added.

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.

8.3.5.3 Dimensionality reduction (Variable reduction):

The earlier methods, namely subset selection and shrinkage, control variance either by using a subset ofthe original variables or by shrinking their coefficients towards zero. In dimensionality reduction, predictors (*X*) are transformed, and the model is set up using the transformed variables after dimensionality reduction.

The number of variables is reduced using the dimensionality reduction method. Principal component analysis is one of the most important dimensionality (variable) reduction techniques.

8.3.6 Polynomial Regression Model:

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

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

In the above equation, c_0, c_1, c_2, and c_3 are thecoefficients.

Example: Let us use the below data set of (*X*, *Y*) fordegree 3 polynomial.

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

As you can observe, the regression line (refer to Fig. 8.16) is slightly curved for polynomial degree 3 with the above 15 data points. The regression line will curve further if we increase the polynomial degree (refer to Fig. 8.17). At the extreme value as shown below, the regression line will be overfitting into all the original values of *X*.

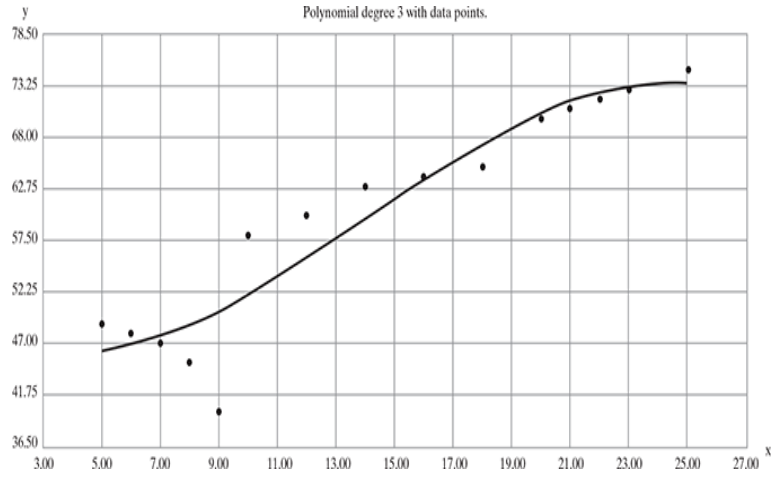


FIG. 8.16 Polynomial regression degree 3

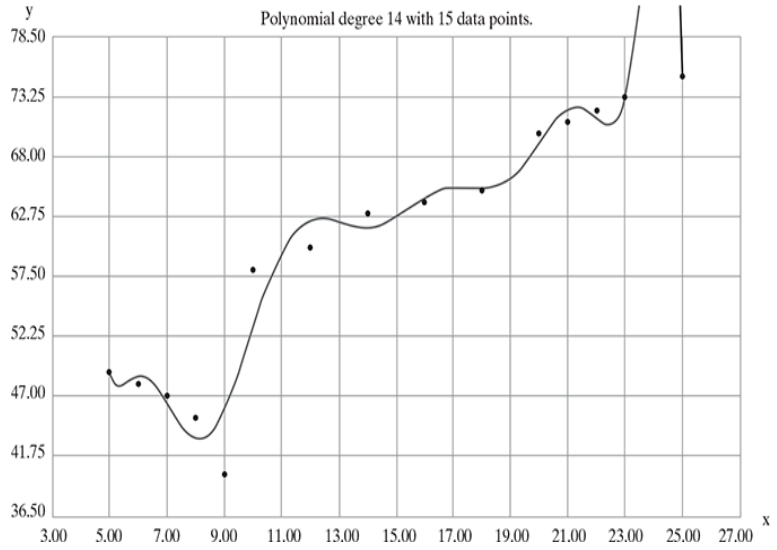
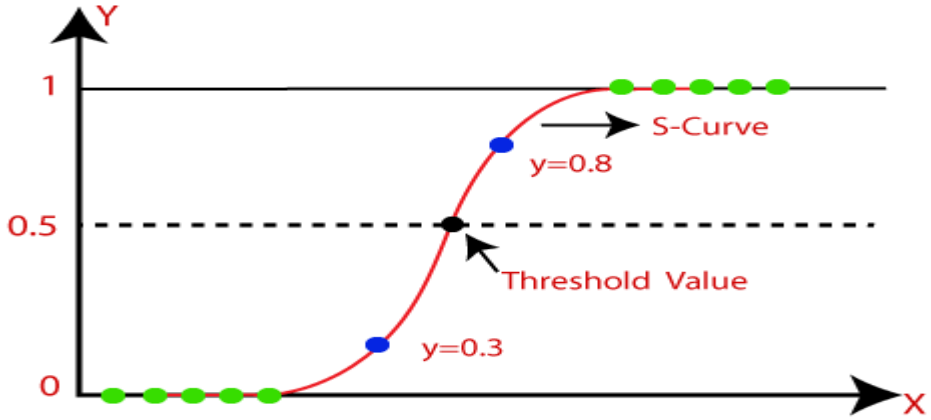


FIG. 8.17 Polynomial regression degree 14

8.3.7 Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function or Sigmoid function, which predicts two maximum values (0 or 1).



Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Logistic Regression Equation:

- The Logistic regression equation can be obtained from the Linear Regression equation.
- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Type of Logistic Regression:

- On the basis of the categories, Logistic Regression can be classified into three types:
  1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
  2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
  3. **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Example: Weather Prediction

- Weather predictions are the result of logical regression. Here, we analyse the data of the previous weather reports and predict the possible outcome for a specific day. But logical regression would only predict categorical data, like if its going to rain or not.

Example 2:

To illustrate this, it is convenient to segregate 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, we will get something like

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

An explanation of logistic regression begins with an explanation of the logistic function, which always takes values between zero and one. The logistic formulae are stated in terms of the probability that $Y = 1$, which is referred to as P. The probability that Y is 0 is $1 - P$.

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

$$\ln(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

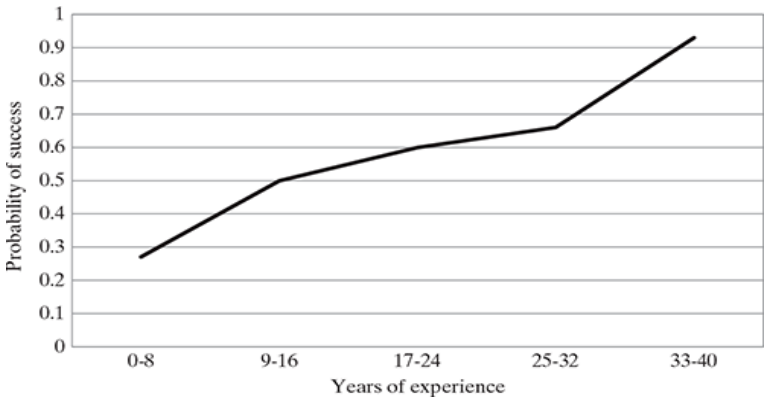


FIG. 8.18 Logistic regression

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. So, if we know the regression equation, we could, theoretically, calculate the expected probability that $Y = 1$ for a given value of X.

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

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

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 $P(\text{male}|\text{height} = 150)$.

$y = e\text{^}(a + b \times X)/(1 + e\text{^}(a + b \times X))$

$$y = \exp(-100 + 0.6 \times 150)/(1 + EXP(-100 + 0.6 \times X) y = 0.000046$$

or a probability of near zero that the person is a male.

Assumptions in logistic regression

The following assumptions must hold when building alogistic regression model:

- There exists a linear relationship between logit 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 '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*]
  - *n* = # of records in the data
  - *p* = probability of success (pass, responder)

8.3.8 Maximum Likelihood Estimation:

The coefficients in a logistic regression are estimated using a process called Maximum Likelihood Estimation (MLE). First, let us understand what is likelihood function before moving to MLE. A fair coin outcome flipsequally 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. Whenever this number (*P*) is greater than 0.5, it is said to be in favour of Head. Whenever *P* is lesser than 0.5, it is said to be against the outcome of getting Head.

Let us represent '*n*' flips of coin as $X_1, X_2, X_3, \ldots, X_n$.
Now X_i can take the value of 1 or 0.

X_i = 1 if Head is the outcome
X_i = 0 if Tail is the outcome

When we use the Bernoulli distribution representseach flip of the coin:

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

Each observation X_i is independent and also identically distributed (iid), and the joint distribution simplifies to a product of distributions.

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

where #*H* is the number of flips that resulted in theexpected outcome (heads in this case).

The likelihood equation is

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

But the likelihood function is not a probability. Thelikelihood for some coins may be 0.25 or 0 or 1.

MLE is about predicting the value for the parametersthat maximizes the likelihood function.

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

SHORT ANSWER-TYPE QUESTIONS (5 MARKS EACH)

1. What is a dependent variable and an independent variable in a linear equation?
2. What is simple linear regression? Give one example.
3. Define slope in a linear regression.
4. 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 $(2, 2)$.
5. What are the conditions of a positive slope in linear regression?
6. What are the conditions of a negative slope in linear regression?
7. What is multiple linear regression?
8. Define sum of squares due to error in multiple linear regression.
9. Define sum of squares due to regression in multiple linear regression.
10. What is multicollinearity in regression equation?
11. What is heteroskedasticity?
12. Explain ridge regression.
13. Explain lasso regression.
14. What is polynomial regression?
15. Explain basis function.
16. Explain logistic regression.

LONG ANSWER-TYPE QUESTIONS (10 MARKS EACH)

1. Define simple linear regression using a graph explaining slope and intercept.
2. Explain rise, run, and slope in a graph.
3. Explain slope, linear positive slope, and linear negative slope in a graph along with various conditions leading to the slope.
4. Explain curve linear negative slope and curve linear positive slope in a graph.
5. Explain maximum and minimum point of curves through a graph.
6. Explain ordinary least square with formula for a and b.
7. Explain the OLS algorithm with steps.
8. What is standard error of the regression? Draw a graph to represent the same.
9. Explain multiple linear regression with an example.
10. Explain the assumptions in regression analysis and BLUE concept.
11. Explain two main problems in regression analysis.
12. How to improve accuracy of the linear regression model?
13. Explain polynomial regression model in detail with an example.
14. Explain logistic regression in detail.
15. What are the assumptions in logistic regression?
16. Discuss maximum likelihood estimation in detail.