Chapter 8

Supervised Learning: Regression

8.1 INTRODUCTION

OBJECTIVE OF THE CHAPTER:

In the last two chapters, you have got quite a good conceptual overview of supervised learning algorithm for categorical data prediction. You got a detailed understanding of all the popular models of classification that are used by machine learning practitioners to solve a wide array of prediction problems where the target variable is a categorical variable.

In this chapter, we will build concepts on prediction of numerical variables – which is another key area of supervised learning. This area, known as regression, focuses on solving problems such as predicting value of real estate, demand forecast in retail, weather forecast, etc.

First, you will be introduced to the most popular and simplest algorithm, namely simple linear regression. This model roots from the statistical concept of fitting a straight line and the least squares method. We will explore this algorithm in detail. In this same context, we will also explore the concept of multiple linear regression.

We will then briefly touch upon the other important algorithms in regression, namely multivariate adaptive regression splines, logistic regression, and maximum likelihood estimation.

By the end of this chapter, you will gain sufficient knowledge in all the aspects of supervised learning and become ready to start solving problems on your own.

8.2 EXAMPLE OF REGRESSION

We have mentioned many times that real estate price prediction is a problem that can be solved by supervised learning or, more specifically, by regression. So, what this problem really is? Let us delve a little deeper into the problem.

New City is the primary hub of the commercial activities in the country. In the last couple of decades, with increasing globalization, commercial activities have intensified in New City. Together with that, a large number of people have come and settled in the city with a dream to achieve professional growth in their lives. As an obvious fall-out, a large number of housing projects have started in every nook and corner of the city. But the demand for apartments has still outgrown the supply. To get benefit from this boom in real estate business, Karen has started a digital market agency for buying and selling real estates (including apartments, independent houses, town houses, etc.). Initially, when the business was small, she used to interact with buyers and sellers personally and help them arrive at a price quote – either for selling a property (for a seller) or for buying a property (for a buyer). Her long experience in real estate business helped her develop an intuition on what the correct price quote of a property could be – given the value of certain standard parameters such as area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc. However, with the huge surge in the business, she is facing a big challenge. She is not able to manage personal interactions as well as setting the correct price quote for the properties all alone. She hired an assistant for managing customer interactions. But the assistant, being new in the real estate business, is struggling with price quotations. How can Karen solve this problem?

Fortunately, Karen has a friend, Frank, who is a data scientist with in-depth knowledge in machine learning models. Frank comes up with a solution to Karen's problem. He builds a model which can predict the correct value of a real estate if it has certain standard inputs such as area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc. Wow, that sounds to be like Karen herself doing the job! Curious to know what model Frank has used? Yes, you guessed it right. He used a regression model to solve Karen's real estate price prediction problem.

So, we just discussed about one problem which can be solved using regression. In the same way, a bunch of other problems related to prediction of numerical value can be solved using the regression model. 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 in the context of Karen's problem. This variable is presumed to be functionally related to one (say, X) or more independent variables called predictors. In the context of Karen's problem, Frank used area of the property, location, floor, etc. as predictors of the model that he built. In other words, the dependent variable depends on independent variable(s) or predictor(s). Regression is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variable(s) (X), i.e. to find the function 'f' for the association $Y = f(X)$.

8.3 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)

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 predictor variable as shown in Figure 8.1.

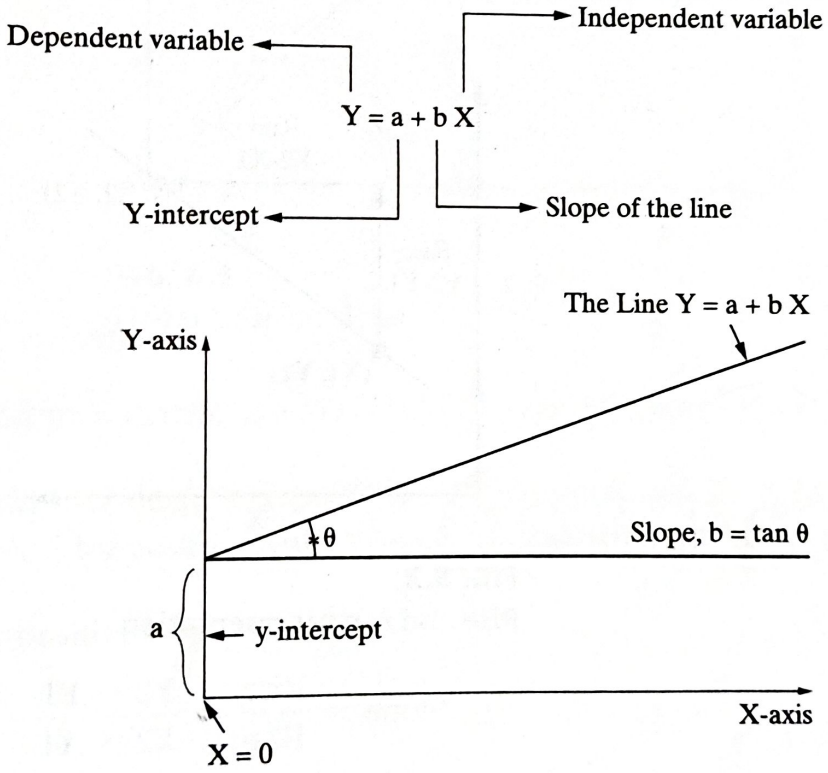

FIG. 8.1
Simple linear regression

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 the model as

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

where 'a' and 'b' are intercept and slope of the straight line, 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 Fig. 8.1).

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 as given below:

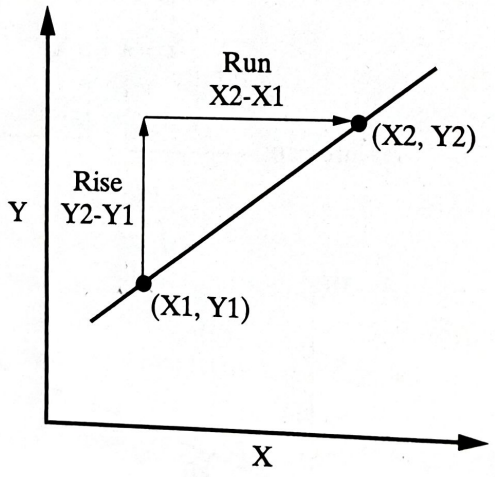

FIG. 8.2
Rise and run representation

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

Example of slope

Let us find the slope of the graph where the lower point on 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

- 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 from left to right (refer to Fig. 8.3).

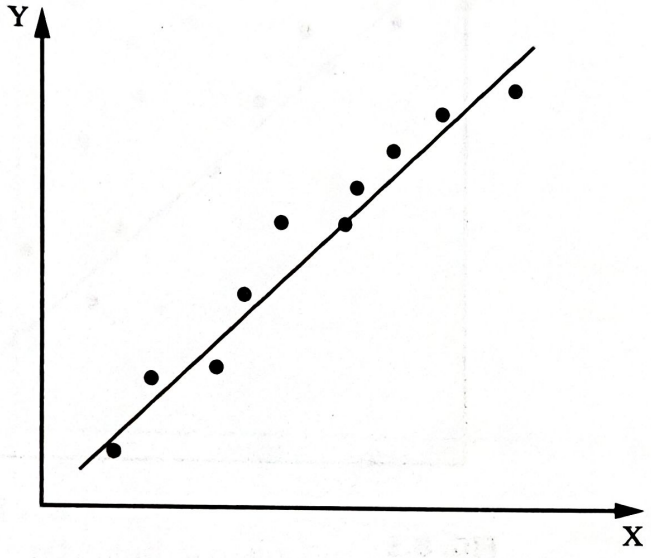

FIG. 8.3
Linear positive slope

$$\text{Slope} = \text{Rise/Run} = (Y_2 - Y_1) / (X_2 - X_1) = \text{Delta}\ (Y) / \text{Delta}(X)$$

- 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

Curve linear positive slope

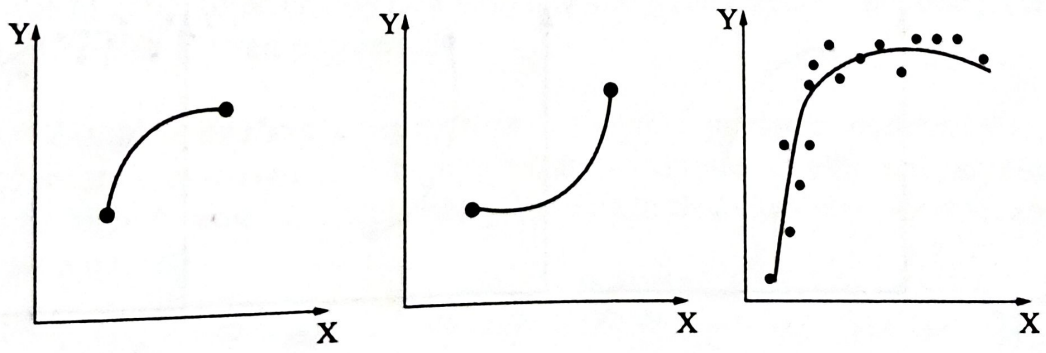

FIG. 8.4
Curve linear positive slope

Curves in these graphs (refer to Fig. 8.4) slope upward from left to right.

$$\text{Slope} = (Y_2 - Y_1) / (X_2 - X_1) = \text{Delta}\ (Y) / \text{Delta}(X)$$

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 (refer to Fig. 8.5).

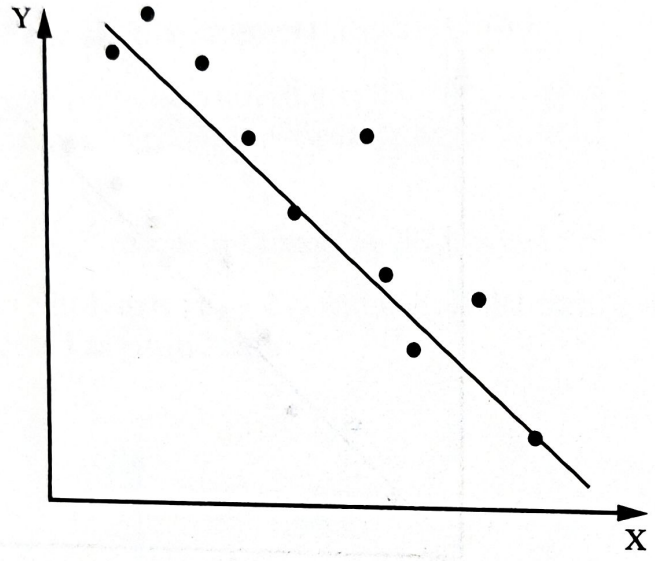

FIG. 8.5
Linear negative slope

$$\text{Slope} = \text{Rise/Run} = (Y_2 - Y_1) / (X_2 - X_1) = \text{Delta}\ (Y) / \text{Delta}(X)$$

- 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

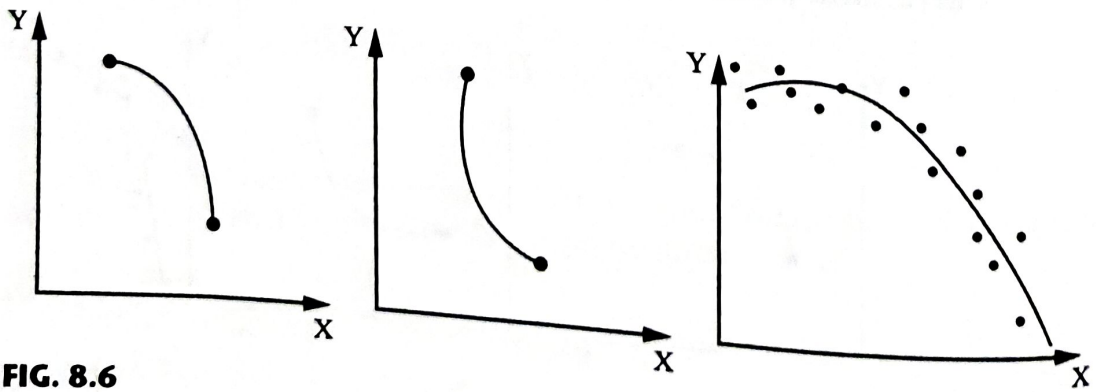

FIG. 8.6
Curve linear negative slope

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

$$\text{Slope} = (Y_2 - Y_1) / (X_2 - X_1) = \text{Delta}\ (Y) / \text{Delta}(X)$$

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.

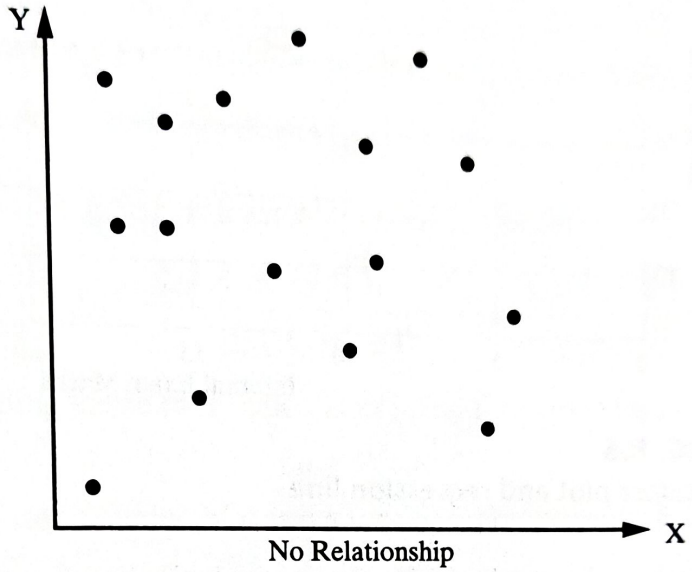

FIG. 8.7
No relationship graph

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.

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

Now that we have some context of the simple regression model, let us try to explore an example to understand clearly how to decide the parameters of the model (i.e. values of a and b) for a given problem.

8.3.1.4 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

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.

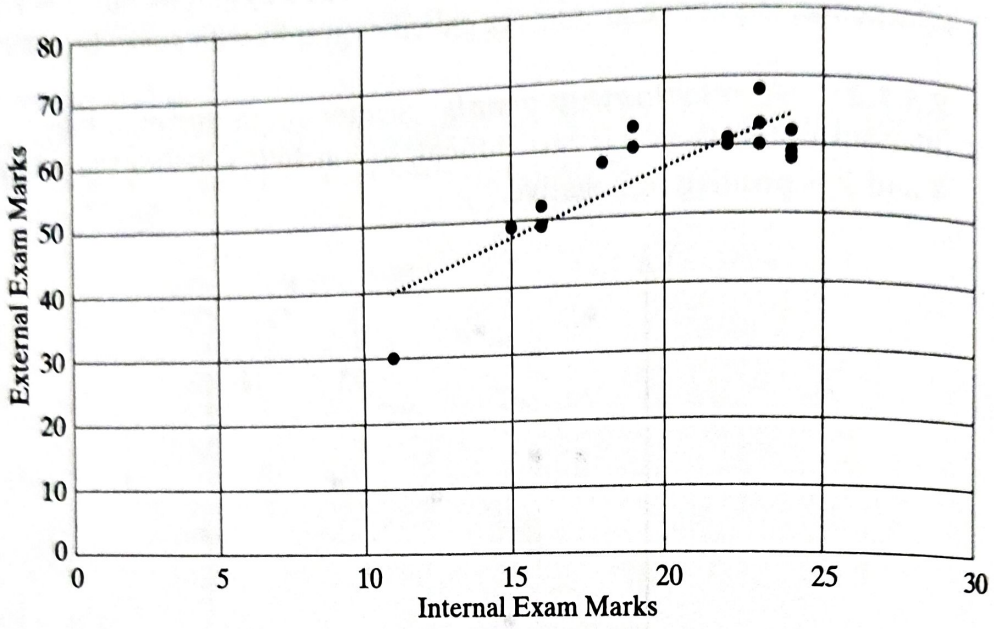

FIG. 8.8
Scatter plot and regression line

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.

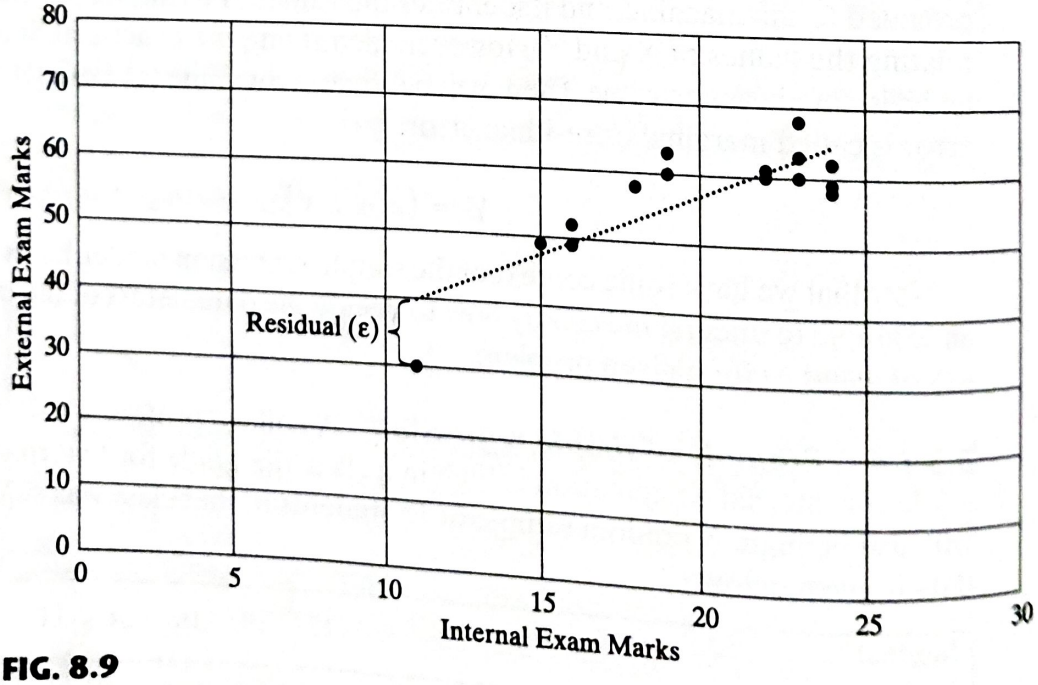

FIG. 8.9
Residual error

As we know, in simple linear regression, the line is drawn 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 (SSE) $\left(\text{i.e. } \sum_i \varepsilon_i^2\right)$.

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

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

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

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

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 $X = 299$
Sum of $Y = 852$
Mean X, $M_X = 19.93$
Mean Y, $M_Y = 56.8$
Sum of squares (SS_X) = 226.9333
Sum 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, $M_{\text{Ext}} = 19.04 + 1.89 \times M_{\text{Int}}$

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	[illegible]	-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	[illegible]
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

FIG. 8.10
Detailed calculation of regression parameters

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)

Interpretation of the intercept

As we have already seen, the simple linear regression model built on the data in the example is

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

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

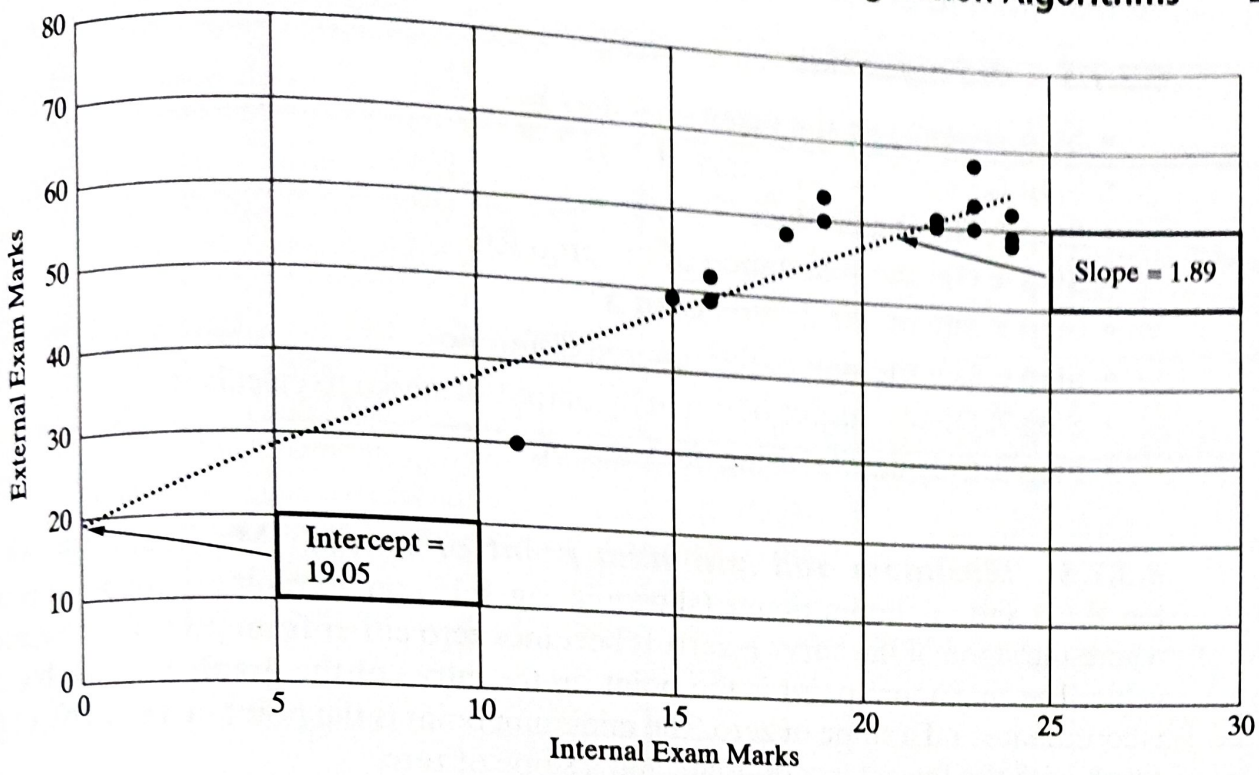

FIG. 8.11
Extended version of the regression graph

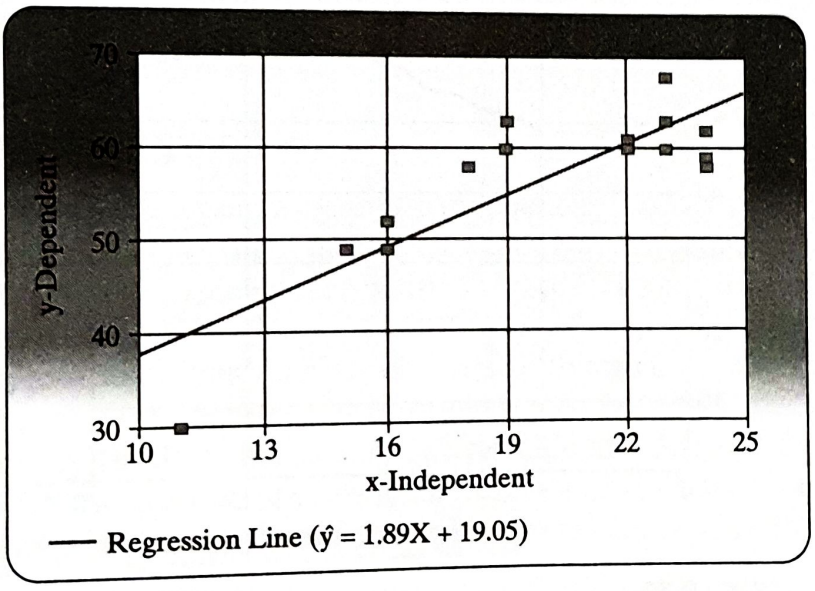

FIG. 8.12
Zoom-in regression line

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.

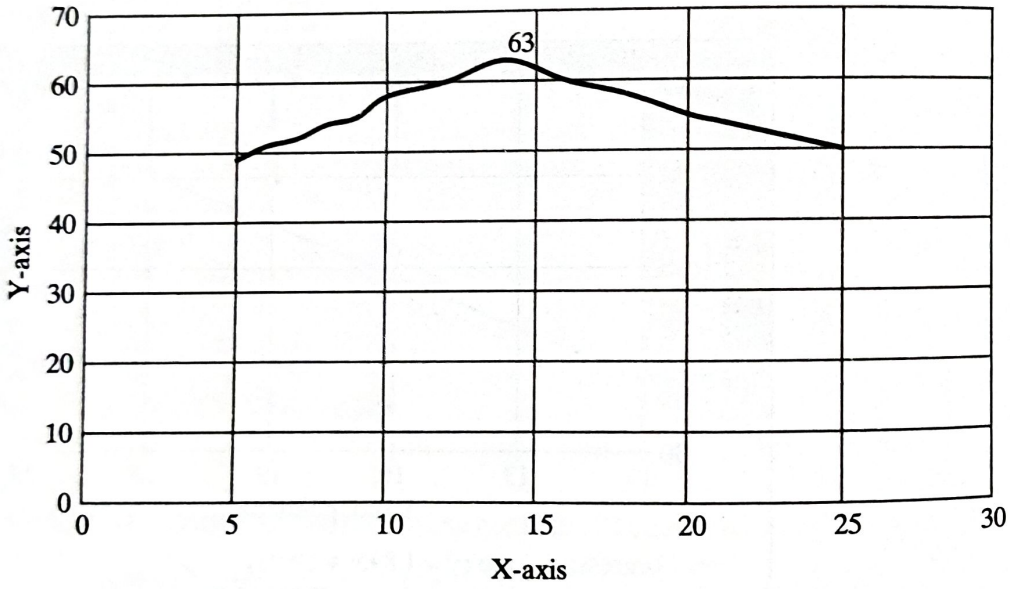

FIG. 8.13
Maximum point of curve

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.

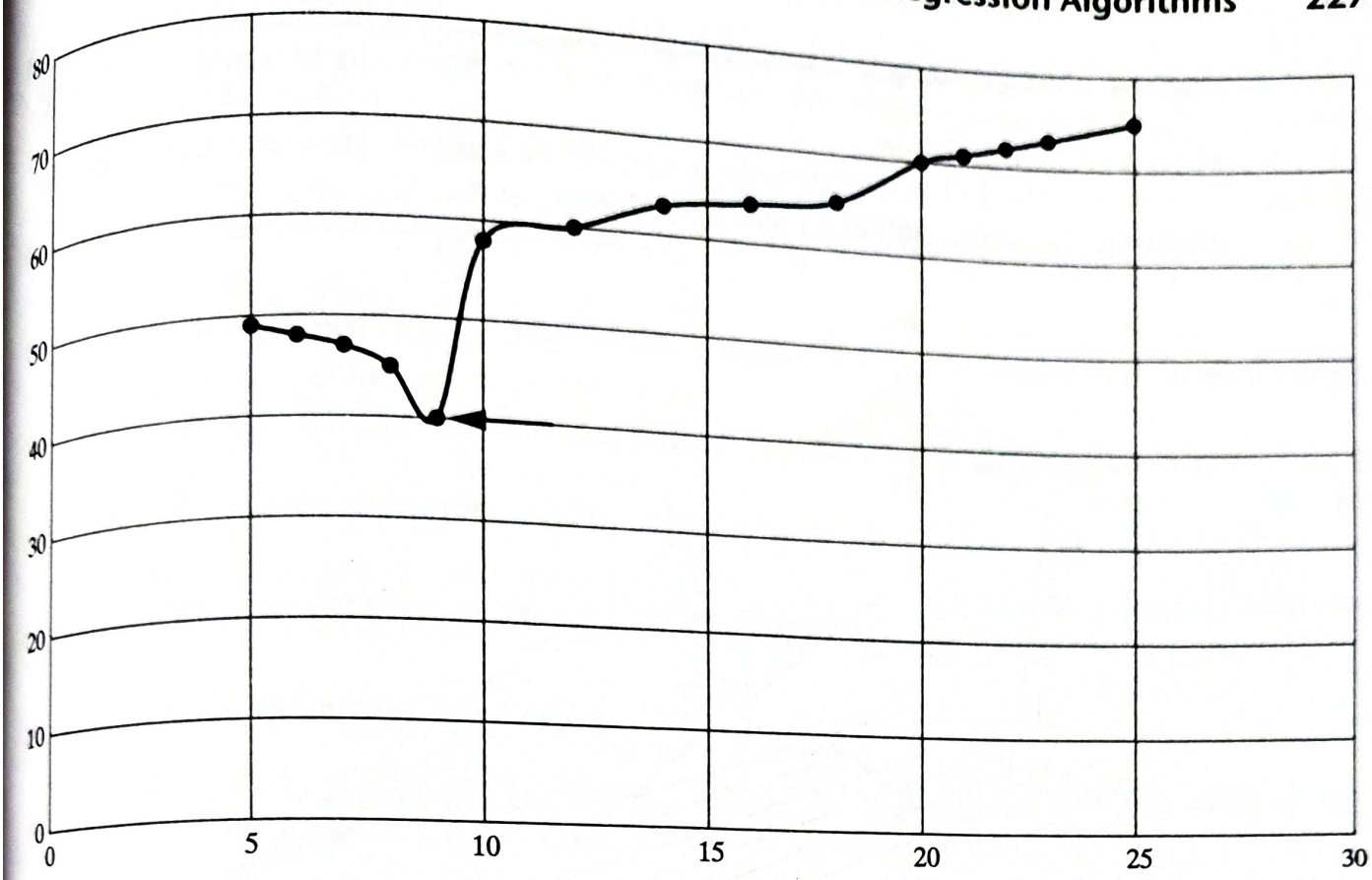

FIG. 8.14
Minimum point of curve

8.3.2 Multiple Linear Regression

In a multiple regression model, two or more independent variables, i.e. predictors are involved in the model. If we think in the context of Karen's problem, in the last section, we came up with a simple linear regression by considering Price of a Property as the dependent variable and the Area of the Property (in sq. m.) as the predictor variable. However, location, floor, number of years since purchase, amenities available, etc. are also important predictors which should not be ignored. Thus, if we consider Price of a Property (in $) as the dependent variable and Area of the Property (in sq. m.), location, floor, number of years since purchase and amenities available as the independent variables, we can form a multiple regression equation as shown below:

$$\text{Price}_{\text{Property}} = f(\text{Area}_{\text{Property}}, \text{location}, \text{floor}, \text{Ageing}, \text{Amenities})$$

The simple linear regression model and the multiple regression model assume that the dependent variable is continuous.

The following expression describes the equation involving the relationship with two predictor variables, namely X_1 and X_2.

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

The model describes a plane in the three-dimensional space of $\hat{Y}$, 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**. Parameter b_1 represents the change in the mean

response corresponding to a unit change in X_1 when X_2 is held constant. Parameter b_2 represents the change in the mean response corresponding to a unit change in X_2 when X_1 is held constant.

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).

$$\hat{Y} = 22 + 0.3X_1 + 1.2X_2$$

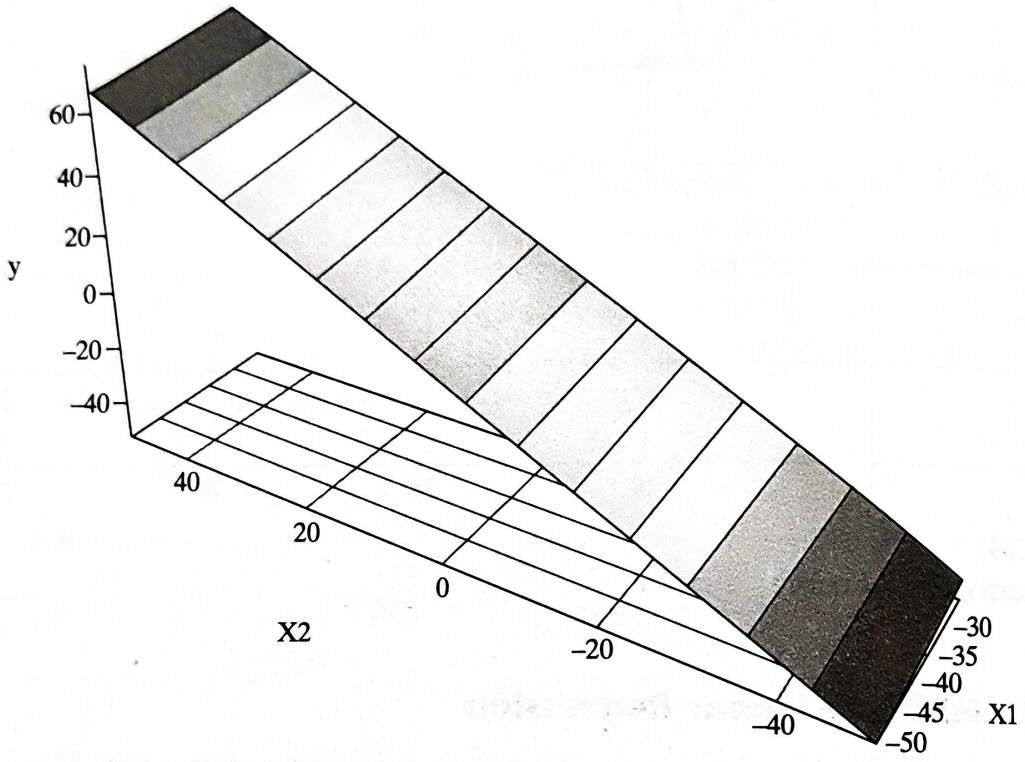

FIG. 8.15
Multiple regression plane

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

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$$

While finding the best fit line, we can fit either a polynomial or curvilinear regression. These are known as polynomial or curvilinear regression, respectively.

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 this is 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 the situation 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. When multicollinearity is present, it increases the standard errors of the coefficients. By overinflating the standard errors, multicollinearity tries to make some variables statistically insignificant when they actually should be significant (with lower standard errors). One way to gauge multicollinearity is to calculate the Variance Inflation Factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if the predictors are correlated. If no factors are correlated, the VIFs will be equal to 1.

The assumption of no perfect collinearity states that there is no exact linear relationship among the independent variables. This assumption implies two aspects of the data on the independent variables. First, none of the independent variables, other than the variable associated with the intercept term, can be a constant. Second, variation in the X's is necessary. In general, the more variation in the independent variables, the better will be the OLS estimates in terms of identifying the impacts of the different independent variables on the dependent variable.

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

Let us understand bias and variance in the regression model before exploring how to improve the same. 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 can try 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 two best-known techniques for shrinking the regression coefficients towards zero are

1. ridge regression
2. lasso (Least Absolute Shrinkage Selector Operator)

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

Minimization objective of ridge = LS Obj + $\alpha \times$ (sum of square 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 the number of variables k is quite large. Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.

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

Minimization objective of ridge = LS Obj + $\alpha \times$ (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:

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

In best subset selection, we fit a separate least squares regression 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 subset selection:

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 too step 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 of the 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 the coefficients.

Example: Let us use the below data set of (X, Y) for degree 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.

In best subset selection, we fit a separate least squares regression 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 subset selection:

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 too step 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 of the 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 the coefficients.

Example: Let us use the below data set of (X, Y) for degree 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.

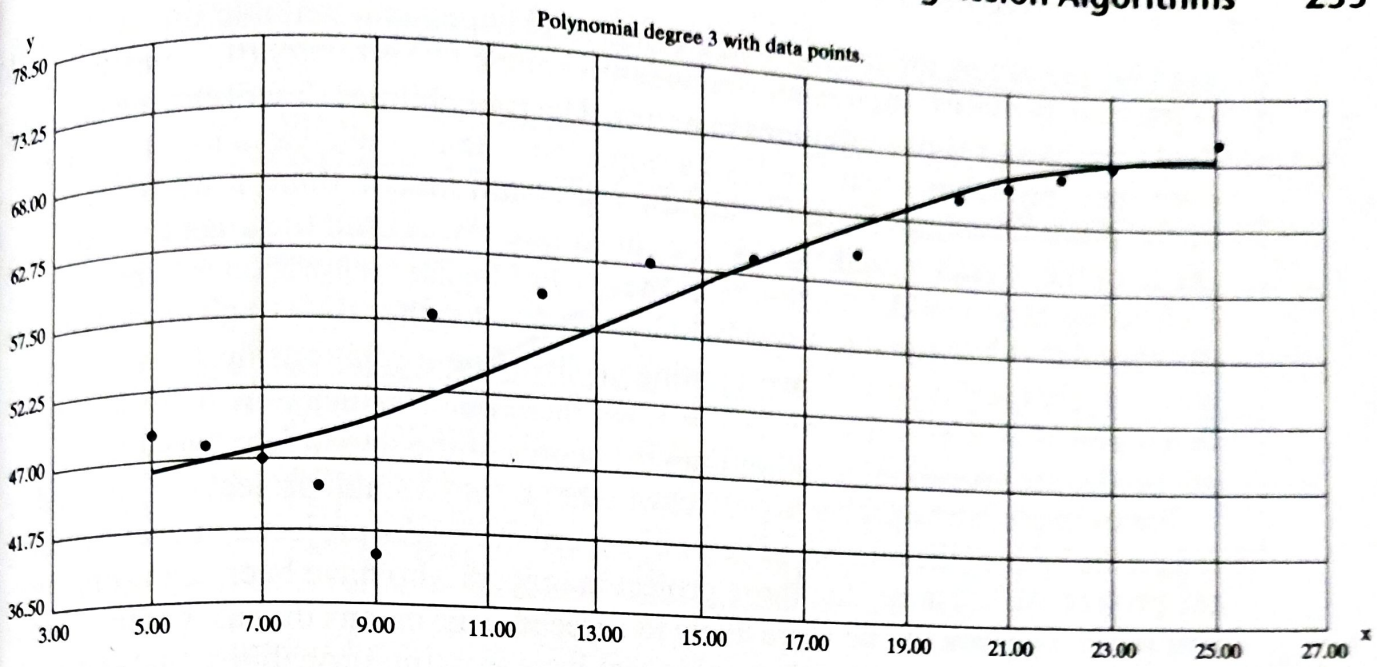

FIG. 8.16
Polynomial regression degree 3

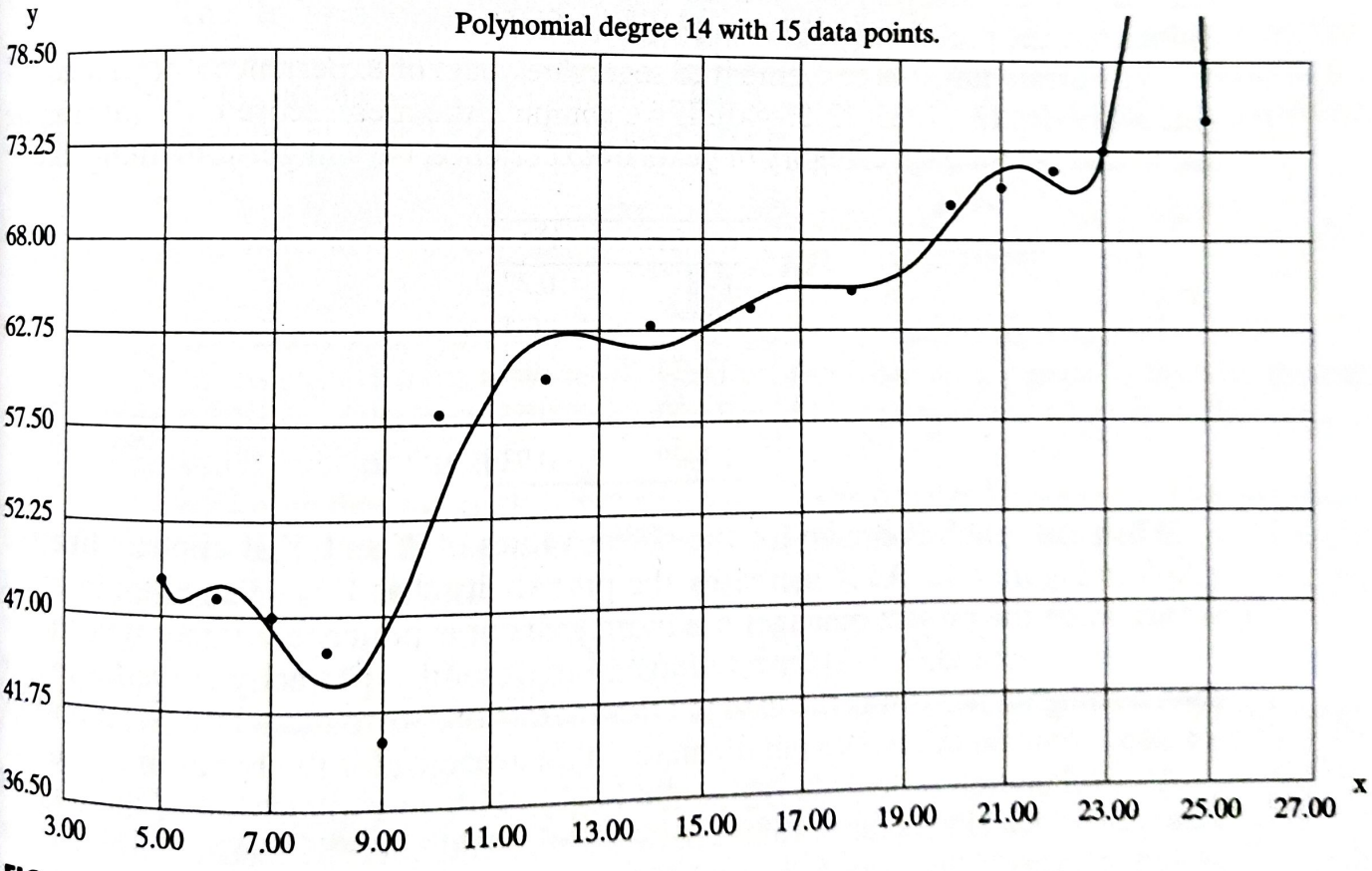

FIG. 8.17
Polynomial regression degree 14

8.3.7 Logistic Regression

Logistic regression is both classification and regression technique depending on the scenario used. Logistic regression (logit regression) is a type of regression analysis

used for predicting the outcome of a categorical dependent variable similar to OLS regression. In logistic regression, dependent variable (Y) is binary (0,1) and independent variables (X) are continuous in nature. The probabilities describing the possible outcomes (probability that $Y = 1$) of a single trial are modelled as a logistic function of the predictor variables. In the logistic regression model, there is no R^2 to gauge the fit of the overall model; however, a chi-square test is used to gauge how well the logistic regression model fits the data. 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. That is, if X and Y have a strong positive linear relationship, the probability that a person will have a score of $Y = 1$ will increase as values of X increase. So, we are predicting probabilities rather than the scores of the dependent variable.

For example, we might try to predict whether or not a small project will succeed or fail on the basis of the number of years of experience of the project manager handling the project. We presume that those 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. If we take a hypothetical example in which 60 already executed projects were studied and the years of experience of project managers ranges from 0 to 20 years, we could represent this tendency to increase the probability that $Y = 1$ with a graph.

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

When the graph is drawn for the above values of X and Y, it appears like the graph in Figure 8.18. As X increases, the probability that $Y = 1$ increases. In other words, when the project manager has more years of experience, a larger percentage of projects succeed. 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.

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 $1 - P$.

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

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

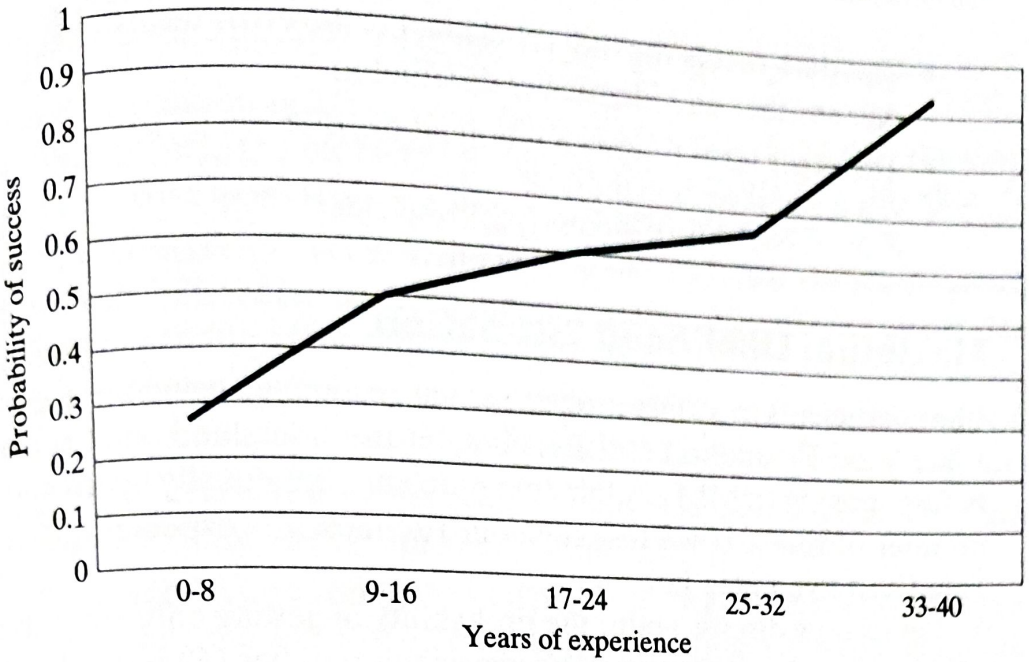

FIG. 8.18
Logistic regression

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

$$y = e\text{^}(a + b \times X)/(1 + e\text{^}(a + b \times X))$$
$$\text{y} = \exp(-100 + 0.6 \times 150)/(1 + \text{EXP}(-100 + 0.6 \times X)$$
$$y = 0.000046$$

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

Assumptions in logistic regression

The following assumptions must hold when building a logistic 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 flips 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. 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 represents each 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.

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

where $\#H$ is the number of flips that resulted in the expected outcome (heads in this case).

The likelihood equation is

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

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.

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

8.4 SUMMARY

- In supervised learning, when we are trying to predict a real-value variable such as 'Price', 'Weight', etc., the problem falls under the category of regression. A regression problem tries to forecast results as a continuous output.
- Dependent Variable (Y) is the value to be predicted. This variable is presumed to be functionally related to the independent variable (X). In other words, dependent variable(s) depends on independent variable(s).
- Independent Variable (X) is called as predictor. The independent variable (X) is used in a regression model to estimate the value of the dependent variable (Y).
- Regression is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variables (X).
- If the regression involves only one independent variable, it is called simple regression. Thus, if we take 'Price of a used car' as the dependent variable and the 'Year of manufacturing of the car' as the independent variable, we can build a simple regression.
- Slope represents how much the line in a graph changes in the vertical direction (Y-axis) over a change in the horizontal direction (X-axis). Slope is also referred as the rate of change in a graph.
- Maximum and minimum points on a graph are found at points where the slope of the curve is zero. It becomes zero either from positive or from negative value.
- If two or more independent variables are involved, it is called multiple regression. Thus, if we take 'Price of a used car' as the dependent variable and year of manufacturing (Year), brand of the car (Brand), and mileage run (Miles run) as the independent variables, we can form a multiple regression equation as given below:

  Price of a used car ($) = function (Year, Brand, Miles run)

- Multicollinearity is the situation in which the degree of correlation is not only between the dependent variable and the independent variable, but there also exists a strong correlation within (among) the independent variables itself.
- 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, normally (identically) distributed (iid). The error terms should not be related to each other.
- Accuracy of linear regression can be improved using the following three methods:

  1. Shrinkage Approach
  2. Subset Selection
  3. Dimensionality Reduction

- 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.

- In logistic regression, the dependent variable (Y) is binary (0,1) and independent variables (X) are continuous in nature. The probabilities describing the possible outcomes (probability that $Y = 1$) of a single trial are modelled as a function of the explanatory (predictor) variables by using a logistic function.

SAMPLE QUESTIONS

MULTIPLE-CHOICE QUESTIONS

1. When we are trying to predict a real-value variable such as '$', 'Weight', the problem falls under the category of
   (a) Unsupervised learning
   (b) Supervised regression problem
   (c) Supervised classification problem
   (d) Categorical attribute

2. Price prediction of crude oil is an example of
   (a) Unsupervised learning
   (b) Supervised regression problem
   (c) Supervised classification problem
   (d) Categorical attribute

3. Value to be predicted in machine learning is called as
   (a) Slope
   (b) Regression
   (c) Independent variable
   (d) Dependent variable

4. This is called as predictor.
   (a) Slope
   (b) Regression
   (c) Independent variable
   (d) Dependent variable

5. This is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variables (X).
   (a) Slope
   (b) Regression
   (c) Classification
   (d) Categorization

6. If the regression involves only one independent variable, it is called as
   (a) Multiple regression
   (b) One regression
   (c) Simple regression
   (d) Independent regression

7. Which equation represents simple imperfect relationship?
   (a) $Y = (a + bx) + \varepsilon$
   (b) $Y = (a + bx)$
   (c) DY = Change in Y / Change in X
   (d) $Y = a + b1X1 + b2X2$

8. Which equation represents simple perfect relationship?
   (a) $Y = (a + bx) + \varepsilon$
   (b) $Y = (a + bx)$
   (c) DY = Change in Y / Change in X
   (d) $Y = a + b1X1 + b2X2$

9. What is the formula for slope in a simple linear equation?
(a) $Y = (a + bx) + \epsilon$
(b) $Y = (a + bx)$
(c) DY = Change in Y / Change in X
(d) $Y = a + b1X1 + b2X2$

10. This slope always moves upward on a graph from left to right.
(a) Multilinear slope
(b) No relationship slope
(c) Negative slope
(d) Positive slope

11. This slope always moves downwards on a graph from left to right.
(a) Multilinear slope
(b) No relationship slope
(c) Negative slope
(d) Positive slope

12. Maximum and minimum points on a graph are found at points where the slope of the curve is
(a) Zero
(b) One
(c) 0.5
(d) Random number

13. In the OLS algorithm, the first step is
(a) Calculate the mean of Y and X
(b) Calculate the errors of X and Y
(c) Get the product (multiply)
(d) Sum the products

14. In the OLS algorithm, the last step is
(a) Calculate 'a' using the value of 'b'
(b) Calculate 'b' using the value of 'a'
(c) Get the product (multiply)
(d) Sum the products

15. Which equation below is called as Unexplained Variation?
(a) SSR (Sum of Squares due to Regression)
(b) SSE (Sum of Squares due to Error)
(c) SST (Sum of Squares Total):
(d) R-square (R2)

16. Which equation below is called as Explained Variation?
(a) SSR (Sum of Squares due to Regression)
(b) SSE (Sum of Squares due to Error)
(c) SST (Sum of Squares Total):
(d) R-square (R2)

17. When new predictors (X) are added to the multiple linear regression model, how does R^2 behave?
(a) Decreasing
(b) Increasing or decreasing
(c) Increasing and decreasing
(d) Increasing or remains constant

18. Predicting stochastic events precisely is not possible.
(a) True
(b) False