
UNIT 8 – REGRESSION MODELS

8.0 Introduction

8.1 Expected Learning Outcomes

8.2 Introduction to Regression

8.3 Regression Models

8.3.1 Linear Regression

8.3.2 Multiple Regression

8.3.3 Polynomial Regression

8.3.4 Ridge and Lasso Regression

8.3.5 Support Vector Regression

8.4 Summary

8.5 Answers

8.6 References/Further Readings

8.0 INTRODUCTION

Regression analysis is a key method in statistics that allows us to make predictions and understand how different factors and variables are connected. Currently, where decisions are often based on analyzing large amounts of data—learning about regression models is a crucial skill for anyone interested in data science or analytics for better decision-making.

Regression analysis models help the learner to figure out how one thing changes based on others. For example, a company might want to predict next month's sales based on how much they spend on advertising and the number of sales calls made. These influencing factors, like advertising spending and sales calls, are called predictor variables. The thing you are trying to predict, such as sales, is known as the dependent variable. Similarly, a student could use regression to see how their study hours and attendance affect their test scores, or a homebuyer might want to estimate house prices based on location and size.

Regression models are used everywhere—from businesses forecasting revenue, to doctors estimating disease risks, and to engineers predicting material performance. By using regression, you can make informed guesses about future outcomes and learn which factors matter most.

There are several types of regression models. The simplest, linear regression, looks for a straight-line relationship between variables. Sometimes, relationships are more complex; that's where models like polynomial regression or support vector regression come in, as they can capture curved or intricate patterns in the data.

As you build more advanced models, you might encounter problems like overfitting. Overfitting happens when a model matches the training data too closely and fails to predict new data accurately. This is like memorizing answers for a test rather than understanding the material—great for the practice test, but not for new questions. To help prevent overfitting, special techniques such as Ridge and Lasso regression are used. These techniques add

penalties to the model to keep it from becoming too complex, encouraging simpler, more reliable predictions.

Learning about regression isn't just about making predictions, it's also about measuring how strong and meaningful the relationships are between different things and understanding what really drives outcomes. Throughout this section, you'll discover practical ways to build, apply, and improve regression models, making them powerful tools for solving real-world problems with data, whether you're curious about exam results, sales forecasts, or the price of your dream home.

8.1 EXPECTED LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Understand the fundamental concepts of regression analysis and its role in predictive data analysis
- Distinguish between different types of regression models and their appropriate applications
- Evaluate regression model performance using appropriate metrics and diagnostic tools
- Select appropriate regression techniques based on data characteristics and problem requirements
- Interpret regression results in the context of business and scientific applications

8.2 INTRODUCTION TO REGRESSION

In Unit 7 of this course, we had a brief introduction of Regression and its types, now in this unit we are going to discuss it in detail. Actually, Regression is a statistical method that examines the relationship between a dependent variable (also called response variable, target variable, or outcome variable) and one or more independent variables (also called predictor variables, explanatory variables, or features). The primary goal of regression analysis is to model this relationship in a way that allows us to predict the value of the dependent variable based on known values of the independent variables.

The concept of regression was first introduced by Sir Francis Galton in the 1880s while studying the relationship between the heights of parents and their children. Galton observed that extremely tall parents tend to have children who are tall but closer to the average height - a phenomenon he called "regression toward the mean". This foundational observation led to the development of what we now know as regression analysis.

The mathematical framework for regression was further developed by Karl Pearson and later refined by Ronald Fisher, who introduced the method of maximum likelihood estimation. The evolution continued with the advent of computer technology, enabling the analysis of complex datasets and the development of sophisticated regression techniques.

We learned that Regression is one of the most important techniques in data analysis and predictive modelling because it helps us understand how one variable is influenced by one or more other variables. In simple terms, regression is used when we want to predict or explain the value of an outcome on the basis of certain input factors. It is widely used in research, business, economics, healthcare, engineering, and many other fields because it not only helps in prediction but also explains the nature of relationships among variables

Core Concepts in Regression involves Dependent and Independent Variables, let's understand them.

- The **Dependent variable (Y)** is the variable that we want to predict, estimate, or explain. It is called dependent because its value is assumed to depend on other variables included in the model.
- On the other hand, the **independent variables (X)** are the predictor or explanatory variables i.e. The variables used to predict the dependent variable. These are the input features that influence the outcome

These are the input features that influence or help explain changes in the dependent variable. For example, if we want to predict house prices, then the house price becomes the dependent variable, while factors such as area, number of rooms, and location become the independent variables.

The connection between the Dependent and Independent variable is established by the Regression Equation, the general form of a regression equation can be expressed as:

$$Y = f(X_1, X_2, \dots, X_n) + \varepsilon$$

Where:

- Y is the dependent variable
- $X_1, X_2, \dots, X_n$ are independent variables
- $f()$ is the function that describes the relationship
- ε (epsilon) is the error term representing unexplained variation

Based on the analysis of the Regression equation, the Regression analysis can capture different kinds of relationships between dependent and independent variables, they are as follows:

- In a **linear relationship**, the association between variables can be represented by a straight line, meaning that a change in the predictor leads to a proportional change in the outcome.
- In a **non-linear relationship**, the association follows a curved or more complex pattern, which means the effect of a predictor may change at different levels. Regression can also identify the **direction** of relationships.
- In a **positive relationship**, as one variable increases, the other also tends to increase.
- In a **negative relationship**, as one variable increases, the other tends to decrease.

Understanding the type of relationship is important because it guides the choice of the correct regression model.

This tool of Regression analysis has applications across different domains be it finance, or healthcare or marketing, engineering etc. A few of them are listed below:

- In **finance**, it is used to predict stock prices, assess credit risk, and support portfolio optimization.
- In **healthcare**, regression helps model disease progression, predict treatment outcomes, and analyze factors affecting patient recovery.
- In **marketing**, it is commonly applied to forecast sales on the basis of advertising expenditure, segment customers, and optimize pricing strategies.
- In **engineering**, regression is useful for quality control, performance improvement, and reliability analysis.
- In **economics**, it supports the analysis of factors affecting GDP, inflation, and consumer demand. These applications show that regression is not only a statistical tool but also a practical method for solving real-world problems.

The predictive analysis is one of the key analyses practiced in almost every field of work, and regression plays a vital role in performing predictive analysis. So, let's discuss the role of Regression in Predictive Analytics. Within predictive analytics, regression plays a foundational role due to following reasons:

- It transforms historical data into meaningful predictions about future outcomes. It helps organizations make **data-driven predictions** by using past observations to estimate what may happen next.
- It helps in **identifying key drivers**, meaning it reveals which variables have the strongest influence on the target variable.
- Also, it **quantifies relationships**, as it measures both the strength and the direction of influence between variables. This makes it extremely useful for **decision making**, since managers and researchers can rely on quantitative evidence rather than intuition alone.
- It also contributes to **risk assessment** by helping evaluate possible outcomes and their likelihood, which is valuable in uncertain environments.

In this course we learned that there are a variety of analytical techniques be it Descriptive, or Diagnostic or Prescriptive or Predictive analysis where Regression is one of the most important techniques for data analysis. So, we need to compare it with the other analytical techniques i.e. Regression should also be clearly distinguished from other analytical methods like classification, correlation, time series analysis etc.; the key observations are as follows:

- In **regression vs. classification**, the key difference is that regression predicts **continuous numerical values**, such as income, sales, or temperature, whereas

classification predicts **discrete categories**, such as yes/no, pass/fail, or disease/no disease.

- In **regression vs. correlation**, correlation only measures the strength and direction of association between variables, while regression goes a step further by modelling the predictive relationship and estimating how one variable changes with another.
- In **regression vs. time series analysis**, both methods may be used for prediction, but time series analysis is specifically focused on data observed over time and gives special importance to temporal patterns, trends, and seasonality.

Although regression is a powerful method, it also comes with certain **challenges and limitations**, they are discussed below:

- One common issue is assumption violations, because regression models rely on assumptions such as linearity, independence, homoscedasticity, and normality, which may not always hold in real data.
- Another challenge is overfitting, where a model performs very well on training data but fails to generalize to new or unseen data.
- Multicollinearity is another important problem, occurring when predictor variables are highly correlated with each other, which can make model estimates unstable.
- Outliers can also create difficulties, as extreme values may disproportionately influence the regression line and distort the results.
- Finally, missing data is a common practical issue, since incomplete datasets need careful handling to avoid biased or invalid conclusions. Therefore, while regression is highly useful, it must be applied thoughtfully and with proper diagnostic checks.

☛ Check Your Progress 1

Q1. What is Regression Analysis? Explain its purpose in predictive data analysis.

.....

Q2. Explain the difference between dependent and independent variables with examples.

.....

Q3. Write the general form of a regression equation and explain its components.

.....

Q4. Discuss the advantages and limitations of regression analysis.

.....

Q5. What is the role of regression in predictive analytics?

.....

8.3 REGRESSION MODELS

Regression models form one of the most important components of predictive data analysis because they help in understanding, explaining, and predicting the relationship between a dependent variable and one or more independent variables. In predictive analytics, the

purpose of regression is not limited to estimating future values; it is also used to identify influential factors, measure the strength of relationships, and support evidence-based decision making. Different regression models are used depending on the nature of the data, the type of dependent variable, the relationship among variables, and the assumptions that can reasonably be satisfied. Therefore, a proper understanding of regression models is essential for selecting the right analytical method in real-world applications.

We learned that a regression model is a mathematical representation that expresses how the dependent variable changes with changes in the independent variables. It provides a systematic way to estimate outcomes and interpret how predictor variables contribute to those outcomes. In predictive data analysis, regression models are especially useful because many practical problems involve forecasting a numerical quantity, estimating probability, or modeling patterns in data. For example, an organization may want to predict sales revenue, a bank may want to estimate credit risk, a hospital may want to analyze the impact of patient characteristics on recovery time, or an economist may want to study how inflation affects consumer demand. In all such cases, regression models provide the analytical structure required to move from raw data to meaningful prediction and interpretation.

Regression models are important for several reasons. They help transform historical data into future predictions, reveal hidden relationships among variables, and quantify the magnitude and direction of influence. They also support decision-making by allowing analysts to test hypotheses and compare the effect of different factors. Because of this dual role of explanation and prediction, regression models are widely used in business analytics, economics, healthcare, engineering, finance, social sciences, and machine learning.

The general form of a regression model can be written as:

$$Y = f(X_1, X_2, X_3, \dots, X_n) + \epsilon$$

where Y is the dependent variable, $X_1, X_2, X_3, \dots, X_n$ are the independent variables, $f(\cdot)$ represents the functional relationship between predictors and outcome, and ϵ is the error term. The error term captures the part of variation in the dependent variable that is not explained by the predictors included in the model. In some models, the function f is linear, while in others it may be nonlinear or transformed to fit complex relationships. The choice of the model depends on the underlying structure of the problem being studied.

The most commonly used regression techniques in predictive data analysis include simple linear regression, multiple linear regression, polynomial regression, logistic regression, ridge regression, lasso regression, and elastic net regression. Each of these models has a different purpose and is suitable for different types of problems.

8.3.1 SIMPLE LINEAR REGRESSION

Simple linear regression is the most basic form of regression analysis. It is used when there is one dependent variable and one independent variable, and the relationship between them is assumed to be linear. The equation of simple linear regression is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where Y is the dependent variable, X is the independent variable, β_0 is the intercept, β_1 is the slope coefficient, and ϵ is the error term. The intercept represents the value of Y when $X = 0$, while the slope indicates how much Y changes for a one-unit change in X .

So far as the mathematical foundation of Simple Linear Regression is concerned, the relationship between the dependent variable and the independent variable is expressed through the equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

In this equation, β_0 is the intercept, β_1 is the slope of the regression line, and ϵ is the error term. The values of β_0 and β_1 are estimated using the Ordinary Least Squares (OLS) method. The main objective of OLS is to find the line that best fits the observed data by minimizing the sum of squared residuals, where a residual is the difference between the actual value and the predicted value.

Mathematically, this is written as:

$$\text{Minimize } \sum (Y_i - \hat{Y}_i)^2$$

Here, Y_i represents the actual observed value, while $\hat{Y}_i$ represents the predicted value obtained from the regression line.

By applying this minimization principle, the estimates of the slope and intercept are obtained as follows:

The formulas used are:

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

where $\bar{X}$ is the mean of the independent variable values and $\bar{Y}$ is the mean of the dependent variable values. In simple terms, β_1 shows how much the dependent variable changes for a one-unit change in the independent variable, while β_0 gives the expected value of Y when $X = 0$. These formulas help determine the best-fitting straight line for the given dataset.

This model is useful when the relationship between two variables can be represented by a straight line. For example, a researcher may study the relationship between hours of study and exam marks. If more hours of study generally lead to higher marks, then simple linear regression can estimate that relationship and predict marks for a given study time. Its main strength lies in its simplicity and interpretability, but it becomes insufficient when more than one predictor influences the outcome.

Let's understand the above learned concepts with the help of an Example

Example 1: Using Regression analysis, a teacher wants to study the relationship between **hours studied** and **marks obtained** by students. Here Hours Studied is the **independent variable (X)** and Marks Obtained is the **Dependent variable (Y)**. The data is:

Student	Hours Studied (X)	Marks Obtained (Y)
1	1	52
2	2	55
3	3	61
4	4	66
5	5	70

Solution:

Step 1: Find the Mean of X and Y

Mean of X: $\bar{X} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$ and **Mean of Y:** $\bar{Y} = \frac{52+55+61+66+70}{5} = \frac{304}{5} = 60.8$

So, $\bar{X} = 3, \bar{Y} = 60.8$

Step 2: Calculate $(X_i - \bar{X})$, $(Y_i - \bar{Y})$, $(X_i - \bar{X})(Y_i - \bar{Y})$, and $(X_i - \bar{X})^2$.

X	Y	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	52	-2	-8.8	17.6	4
2	55	-1	-5.8	5.8	1
3	61	0	0.2	0.0	0
4	66	1	5.2	5.2	1
5	70	2	9.2	18.4	4

Now find the totals:

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 17.6 + 5.8 + 0 + 5.2 + 18.4 = 47$$

$$\sum(X_i - \bar{X})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

Step 3: Calculate the Slope β_1 : Using the formula: $\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

Substituting the values we get $\beta_1 = \frac{47}{10} = 4.7$ i.e. the slope is: $\beta_1 = 4.7$

Step 4: Calculate the Intercept β_0 : Using the formula: $\beta_0 = \bar{Y} - \beta_1 \bar{X}$

Substituting the values we get $\beta_0 = 60.8 - (4.7 \times 3) = 60.8 - 14.1 = 46.7$

So, the intercept is: $\beta_0 = 46.7$

Step 5: Write the Regression Equation

Now substitute β_0 and β_1 into the regression equation

$Y = \beta_0 + \beta_1 X$, we get $Y = 46.7 + 4.7X$

This is the **simple linear regression equation** for the given data.

Step 6: Compute Predicted Values Using the equation: $\hat{Y} = 46.7 + 4.7X$

Let us calculate the predicted marks for each value of X .

<u>X</u>	<u>Actual Y</u>	<u>Predicted $\hat{Y} = 46.7 + 4.7X$</u>	<u>Residual $Y - \hat{Y}$</u>
1	52	51.4	0.6
2	55	56.1	-1.1
3	61	60.8	0.2
4	66	65.5	0.5
5	70	70.2	-0.2

Step 7: Check the Residuals and Ordinary Least Square(OLS) method

From the above table we identified the Residual $= Y_i - \hat{Y}_i$

These residuals show the difference between actual marks and predicted marks.

Ordinary Least Square (OLS) method chooses the regression line in such a way that the **sum of squared residuals** becomes minimum. To perform this, tabulate the square of residuals and determine their sum.

X	Residual $Y - \hat{Y}$	Residual Squared $(Y - \hat{Y})^2$
1	0.6	0.36
2	-1.1	1.21

X	Residual $Y - \hat{Y}$	Residual Squared $(Y - \hat{Y})^2$
3	0.2	0.04
4	0.5	0.25
5	-0.2	0.04

$$\sum(Y_i - \hat{Y}_i)^2 = 0.36 + 1.21 + 0.04 + 0.25 + 0.04 = 1.90$$

Since the residuals are small and the sum of squared residuals is low, that is, (1.90), which indicates that the predicted values are close to the observed values, suggesting that the fitted regression line provides a good representation of the data.

This small value indicates that the fitted regression line is close to the actual data points. Because This quantity i.e. $\sum(Y_i - \hat{Y}_i)^2 = 1.90$ measures the total prediction error of the regression line. Each residual is the gap between the actual value and the predicted value. When we square and add these gaps:

- small total $\rightarrow$ actual and predicted values are close
- large total $\rightarrow$ actual and predicted values are far apart

In your example, the residuals were: 0.6, -1.1, 0.2, 0.5, -0.2; and these are all quite small in magnitude, and their squared sum is only 1.90, which suggests the regression line does not deviate much from the observed points.

We observed that the prediction errors are all around less than 1.1 marks, which is very small compared with the scale of marks ranging from 52 to 70. That is why we say the line is close to the actual data points.

Step 8: Interpretation of the Regression Coefficients

Interpretation of Slope $\beta_1 = 4.7$: The slope value of 4.7 means that for every additional 1 hour of study, the marks are expected to increase by 4.7 marks, on average.

So, there is a **positive linear relationship** between study hours and marks.

Interpretation of Intercept $\beta_0 = 46.7$: The intercept value of 46.7 means that if a student studies for 0 hours, the predicted marks would be 46.7.

In practical situations, the intercept may or may not always have direct real-world meaning, but mathematically it helps define the regression line.

Step 9: Prediction Using the Model: Suppose we want to predict the marks of a student who studies for 6 hours. Using the regression equation:

$$\begin{aligned}\hat{Y} &= 46.7 + 4.7(6) \\ \hat{Y} &= 46.7 + 28.2 = 74.9\end{aligned}$$

So, if a student studies for **6 hours**, the predicted marks are: $\hat{Y} = 74.9$.

Thus, the model predicts that the student may score approximately **75 marks**.

Overall Interpretation of Results: The regression equation obtained is $Y = 46.7 + 4.7X$. This equation shows that there is a **clear positive relationship** between hours studied and marks obtained. As the number of study hours increases, the marks also increase. The slope indicates that the increase is about **4.7 marks per extra hour studied**. The actual and predicted values are quite close, and the residuals are small, which suggests that the regression line fits the data well.

From the predictive data analysis point of view, this example shows how simple linear regression can be used both to **understand the relationship** between two variables and to **make future predictions**. It is especially useful when the relationship is approximately linear and only one predictor variable is involved.

This numerical example clearly demonstrates the working of **simple linear regression**. First, the means of the variables were calculated. Then, the slope and intercept were estimated using the **OLS formulas**. After that, the regression equation was formed, predicted values were computed, and the results were interpreted.

Now we extend the understanding of the linear Regression by calculating few more important parameters like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2). By Using the same Example 1, given above, where:

- Actual Y : 52, 55, 61, 66, 70
- Predicted $\hat{Y}$: 51.4, 56.1, 60.8, 65.5, 70.2
- Residuals ($Y - \hat{Y}$): 0.6, -1.1, 0.2, 0.5, -0.2
- Sum of Squared Errors (SSE) = $\sum (Y - \hat{Y})^2 = 1.90$

Calculating Mean Squared Error (MSE): The formula for MSE is $MSE = \frac{\sum (Y_i - \hat{Y}_i)^2}{n}$

Here, $n = 5$. So, $MSE = \frac{1.90}{5} = 0.38$

The interpretation of the resulting **MSE = 0.38** means that the average squared prediction error of the regression model is **0.38**. Since this value is quite low, it indicates that the model's predictions are very close to the actual values. However, because the errors are squared, MSE is expressed in squared units, so it is less directly interpretable than RMSE.

Calculating Root Mean Squared Error (RMSE): The formula for RMSE is

$$RMSE = \sqrt{MSE}$$

$$RMSE = \sqrt{0.38} \approx 0.616$$

The interpretation of the resulting **RMSE ≈ 0.616** means that, on average, the predicted marks differ from the actual marks by about **0.62 marks**. Since the marks in the dataset range roughly from **52 to 70**, an average prediction error of less than 1 mark is very small. This shows that the regression model fits the data very well.

Calculating Coefficient of Determination (R^2): The formula for R^2 is

$$R^2 = 1 - \frac{SSE}{SST}$$

where $SSE = \sum(Y_i - \hat{Y}_i)^2$ and $SST = \sum(Y_i - \bar{Y})^2$,

We already know $\bar{Y} = 60.8$;

Now calculate SST:

Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
52	-8.8	77.44
55	-5.8	33.64
61	0.2	0.04
66	5.2	27.04
70	9.2	84.64

Now, $SST = 77.44 + 33.64 + 0.04 + 27.04 + 84.64 = 222.80$

Already we calculated the Sum of Squared Errors ($SSE = \sum(Y - \hat{Y})^2 = 1.90$)

Now apply the formula: $R^2 = 1 - \frac{SSE}{SST}$

$$R^2 = 1 - \frac{1.90}{222.80}$$

$$R^2 = 1 - 0.00853 \approx 0.9915$$

$$R^2 \approx 0.9915 \text{ i.e. about } \mathbf{99.15\%}$$

The value of $R^2 \approx 0.992$ means that about **99.15% of the variation in marks** is explained by the number of hours studied. This indicates an **excellent fit** of the regression model. In other words, study hours explain almost all the variation in students' marks in this small dataset.

Final Results Summary for the given numerical problem is as follows:

Metric	Value	Interpretation
SSE	1.90	Total squared prediction error is very low
MSE	0.38	Average squared error is low
RMSE	0.616	Average prediction error is about 0.62 marks
R^2	0.9915	About 99.15% of variation in marks is explained by study hours

Overall Interpretation: These values together show that the simple linear regression model provides a very strong fit for the given data.

- The low MSE shows that prediction errors are very small.

- The low RMSE shows that the model's predictions differ from actual marks by less than 1 mark on average.
- The very high R^2 shows that the independent variable, hours studied, explains almost all the variation in the dependent variable, marks obtained.

So, for this dataset, the regression line is highly effective for both understanding the relationship and making predictions.

With the above numerical example, we learned that the Simple linear regression is easy to understand, mathematically elegant, and highly useful in practical situations such as predicting marks from study hours, sales from advertising expense, crop yield from fertilizer use, or income from years of education. It serves as the foundation for more advanced regression techniques used in predictive analytics.

The Linear regression follows some assumptions and they are as follows:

1. Linearity: The relationship between independent and dependent variables is linear
2. Independence: Observations are independent of each other
3. Homoscedasticity: Constant variance of residuals across all levels of independent variables
4. Normality: Residuals are normally distributed
5. No Multicollinearity: Independent variables are not highly correlated (for multiple regression)

Apart from this the Simple Linear Regression has some advantages and Limitations both, they are as follows:

Advantages: Linear regression offers several important advantages, they are:

- It is simple to understand and easy to implement, which makes it one of the most widely used regression techniques.
- It requires relatively low computational effort, so results can be obtained quickly even for moderately sized datasets.
- Another major advantage is that it provides interpretable coefficients, allowing analysts to clearly understand how each independent variable influences the dependent variable.
- It also serves as a strong baseline model for comparing the performance of more advanced predictive techniques.

When the underlying assumptions are reasonably satisfied, linear regression can produce reliable and effective results.

Limitations: Despite its usefulness, linear regression has certain limitations, they are:

- It is based on the assumption that the relationship between variables is linear, which may not always reflect real-world situations.
- The model is also sensitive to outliers, meaning that extreme values can significantly affect the regression line and distort the results.

- In addition, its performance depends on the extent to which key assumptions, such as linearity, homoscedasticity, independence, and normality of errors, are satisfied.
- Another limitation is that linear regression is not well suited for capturing complex or highly non-linear relationships unless additional transformations or advanced extensions are applied.

It is to be noted that the Linear regression is widely applied across many practical domains, a few are listed below:

- In business, it is used for sales forecasting based on advertising expenditure and for revenue prediction based on customer volume.
- In economics, it helps in analysing the relationship between education and income as well as in studying inflation-related trends.
- In the field of sports, linear regression can be used to predict player performance using training and fitness metrics.
- In agriculture, it supports crop yield prediction based on factors such as fertilizer usage and weather conditions.

These applications demonstrate the broad relevance of linear regression as a foundational tool in predictive data analysis.

Check Your Progress 2

Q1. What is a Regression Model? Explain its significance in predictive analytics.

.....

Q2. Define Simple Linear Regression and write its mathematical equation.

.....

Q3. Explain the Ordinary Least Squares (OLS) method used in linear regression.

.....

Q4. Interpret the slope and intercept in a linear regression model..

.....

Q5. List the assumptions and limitations of Simple Linear Regression

.....

.....

8.3.2 MULTIPLE LINEAR REGRESSION

Multiple regression is an extension of simple linear regression in which the dependent variable is explained or predicted using more than one independent variable. This makes it a highly useful technique in predictive data analysis because many real-world outcomes are influenced by several factors simultaneously. Unlike simple linear regression, which studies the effect of only one predictor, multiple regression allows the analyst to model more complex relationships in a more realistic manner. It is especially valuable when the objective is not only to predict the dependent variable accurately but also to understand the separate contribution of each explanatory variable.

The general form of the multiple regression model is expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

In this equation, Y represents the dependent variable, while $X_1, X_2, \dots, X_p$ represent the independent variables. The terms $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the regression coefficients, and ϵ is the error term. The intercept β_0 gives the expected value of Y when all independent variables are equal to zero. Each slope coefficient shows the effect of its corresponding independent variable on the dependent variable, subject to the presence of the other predictors in the model.

For computational convenience and mathematical clarity, multiple regression is often expressed in matrix form as:

$$Y = X\beta + \epsilon$$

Here, Y is an $n \times 1$ vector of dependent variable values, X is an $n \times (p + 1)$ matrix of independent variables including a column of ones for the intercept, β is a $(p + 1) \times 1$ vector of regression coefficients, and ϵ is an $n \times 1$ vector of error terms. Using the Ordinary Least Squares method, the coefficient estimates are obtained from the expression:

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

This matrix formulation provides an efficient way of estimating the regression coefficients, particularly when the model involves many predictors.

Interpretation of Coefficients: In multiple regression, the interpretation of coefficients is based on the principle of *ceteris paribus*, which means “all other things remaining constant.” This means that each coefficient represents the expected change in the dependent variable associated with a one-unit change in the corresponding independent variable, while holding all other independent variables constant. This interpretation is extremely important because, in a multivariable setting, the effect of one predictor is examined after accounting for the influence of the others.

For example, consider the model for predicting house prices:

$$\text{Price} = 20000 + 80 \times \text{Square_Feet} + 15000 \times \text{Bedrooms} + 5000 \times \text{Bathrooms} - 500 \times \text{Age}$$

This equation implies that, holding bedrooms, bathrooms, and age constant, every additional square foot increases the house price by ₹80. Similarly, each additional bedroom increases the price by ₹15,000, each additional bathroom adds ₹5,000, and each additional year of age reduces the price by ₹500. Thus, multiple regression not only predicts house prices but also allows the separate effect of each factor to be clearly understood.

Additional Considerations for Multiple Regression: Although multiple regression is powerful, its application requires careful attention to certain issues, the most important of which is multicollinearity. Multicollinearity occurs when two or more independent variables

are highly correlated with each other. This can create several problems in the model, such as unstable coefficient estimates, inflated standard errors, and difficulty in interpreting the individual contribution of variables. In other words, when predictors overlap substantially in the information they provide, it becomes harder to determine which variable is actually responsible for changes in the dependent variable.

One common method for detecting multicollinearity is the **Variance Inflation Factor (VIF)**. A VIF value greater than 10 is often considered an indication of problematic multicollinearity, although lower thresholds are also sometimes used in practice. When multicollinearity is detected, several remedies are available. Highly correlated variables may be removed from the model, or they may be combined using techniques such as principal component analysis. Another useful solution is the application of regularization methods such as Ridge regression or Lasso regression, which help stabilize coefficient estimates.

Example 2: To understand **multiple regression** in a practical way, let us consider a simple example in which a house price is predicted using two independent variables: **house size** and **number of bedrooms**. Here, the objective is to estimate the effect of both predictors on the dependent variable and also examine whether multicollinearity exists between the independent variables.

Let:

- Y = House Price (in lakh rupees)
- X_1 = House Size (in hundred square feet)
- X_2 = Number of Bedrooms

The observed data are:

Observation	X_1	X_2	Y
1	1	2	16
2	2	3	21
3	3	5	28
4	4	4	31
5	5	6	38

The general form of the multiple regression model is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

For this example, the fitted model will take the form:

$$Y = \beta_0 + \beta_1(\text{House Size}) + \beta_2(\text{Bedrooms}) + \epsilon$$

where:

- β_0 = intercept
- β_1 = coefficient of house size
- β_2 = coefficient of bedrooms
- ϵ = error term

Step 1: Matrix Representation of the Model: In matrix form, multiple regression is written as: $Y = X\beta + \epsilon$

$$\text{where, } Y = \begin{bmatrix} 16 \\ 21 \\ 28 \\ 31 \\ 38 \end{bmatrix}, X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 5 \\ 1 & 4 & 4 \\ 1 & 5 & 6 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

The first column of ones in X is included for the intercept term. The Ordinary Least Squares solution is: $\hat{\beta} = (X'X)^{-1}X'Y$

Step 2: Compute $X'X$ and $X'Y$

First, compute the transpose product $X'X = \begin{bmatrix} 5 & 15 & 20 \\ 15 & 55 & 69 \\ 20 & 69 & 90 \end{bmatrix}$

Next, compute $X'Y = \begin{bmatrix} 134 \\ 456 \\ 587 \end{bmatrix}$

Step 3: Compute $(X'X)^{-1}$:

The inverse of $X'X$ is: $(X'X)^{-1} = \begin{bmatrix} 1.9895 & 0.3158 & -0.6842 \\ 0.3158 & 0.5263 & -0.4737 \\ -0.6842 & -0.4737 & 0.5263 \end{bmatrix}$

Step 4: Estimate the Coefficients: Now apply the formula: $\hat{\beta} = (X'X)^{-1}X'Y$

$$\hat{\beta} = \begin{bmatrix} 1.9895 & 0.3158 & -0.6842 \\ 0.3158 & 0.5263 & -0.4737 \\ -0.6842 & -0.4737 & 0.5263 \end{bmatrix} \begin{bmatrix} 134 \\ 456 \\ 587 \end{bmatrix} = \begin{bmatrix} 8.9579 \\ 4.2632 \\ 1.2632 \end{bmatrix}$$

Thus, the estimated coefficients are: $\beta_0 = 8.9579, \beta_1 = 4.2632, \beta_2 = 1.2632$

Step 5: Write the Estimated Regression Equation:

The fitted multiple regression equation is: $\hat{Y} = 8.9579 + 4.2632X_1 + 1.2632X_2$

This is the required regression model.

Step 6: Compute the Predicted Values: Using the estimated equation:

$$\hat{Y} = 8.9579 + 4.2632X_1 + 1.2632X_2$$

the predicted values are:

Obs.	X_1	X_2	Actual Y	Predicted $\hat{Y}$
1	1	2	16	$8.9579 + 4.2632(1) + 1.2632(2) = 15.75$

Obs.	X_1	X_2	Actual Y	Predicted $\hat{Y}$
2	2	3	21	$8.9579 + 4.2632(2) + 1.2632(3) = 21.26$
3	3	5	28	$8.9579 + 4.2632(3) + 1.2632(5) = 28.05$
4	4	4	31	$8.9579 + 4.2632(4) + 1.2632(4) = 31.06$
5	5	6	38	$8.9579 + 4.2632(5) + 1.2632(6) = 37.85$

So, the predicted values are approximately: 15.75, 21.26, 28.05, 31.06, 37.85

Step 7: Interpretation of the Coefficients:

The estimated model is: $\hat{Y} = 8.9579 + 4.2632X_1 + 1.2632X_2$

The interpretation of the coefficients is as follows:

- **Intercept ($\beta_0 = 8.9579$):** This is the predicted house price when both house size and number of bedrooms are zero. In practice, this may not always have direct physical meaning, but mathematically it is necessary for defining the regression plane.
- **Coefficient of X_1 ($\beta_1 = 4.2632$):** Holding the number of bedrooms constant, a one-unit increase in house size increases the expected house price by **4.2632 lakh rupees**.
- **Coefficient of X_2 ($\beta_2 = 1.2632$):** Holding house size constant, an additional bedroom increases the expected house price by **1.2632 lakh rupees**.

This is the **ceteris paribus** interpretation of multiple regression.

Step 8: Detecting Multicollinearity: One of the major concerns in multiple regression is **multicollinearity**, which arises when the independent variables are highly correlated with each other.

(a) Correlation Between X_1 and X_2 : Using the given data, the correlation between house size and bedrooms is: $r_{X_1X_2} = 0.90$. This indicates a **strong positive correlation** between the two predictors.

(b) Coefficient of Determination for Auxiliary Regression: Since there are only two independent variables, we can regress X_1 on X_2 (or vice versa). In the two-variable case:

$$R_1^2 = r^2 = (0.90)^2 = 0.81$$

Thus, about **81% of the variation in X_1** is explained by X_2 , which suggests considerable overlap between the predictors.

(c) Variance Inflation Factor (VIF): The formula for VIF is: $VIF_j = \frac{1}{1-R_j^2}$

So, $VIF_1 = \frac{1}{1-0.81} = \frac{1}{0.19} = 5.26$; Similarly, $VIF_2 = 5.26$

(d) Tolerance: Tolerance is the reciprocal of VIF:

$$\text{Tolerance} = \frac{1}{VIF} = \frac{1}{5.26} \approx 0.19$$

Interpretation of Multicollinearity Results: The correlation value of **0.90** shows that the two independent variables are strongly related. The VIF value of **5.26** is above 5, which suggests **moderate to fairly high multicollinearity**, although it is still below the more severe cutoff of 10 that is often used in practice. The tolerance value of **0.19** is also relatively low, further confirming that the predictors share substantial common information.

This means that although the regression model can still be estimated, the coefficient estimates may become somewhat unstable if the dataset changes. In practical analysis, such a situation may require closer examination, variable selection, or regularization methods such as Ridge regression.

The fitted model shows that both house size and number of bedrooms positively affect house price. House size has the stronger impact, since its coefficient is much larger than that of bedrooms. This suggests that, in this example, increasing the size of the house contributes more to price than merely increasing the number of bedrooms.

At the same time, the multicollinearity analysis shows that house size and bedrooms are themselves strongly related. This is expected in real-world housing data because larger houses often tend to have more bedrooms. Therefore, while the model is useful for prediction, interpretation of the individual coefficients should be done carefully.

This numerical example demonstrates how **multiple regression** is applied using both its **equation form** and **matrix representation**. The coefficients were estimated through the OLS formula:

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

leading to the regression equation:

$$\hat{Y} = 8.9579 + 4.2632X_1 + 1.2632X_2$$

The example also showed how to compute and interpret **multicollinearity** using correlation, R^2 , VIF, and tolerance. The results indicate that while the model is meaningful and interpretable, the predictors exhibit moderate multicollinearity, which is an important consideration in multiple regression analysis.

Next, we are going to discuss the Model Selection Techniques: In multiple regression, another important issue is the selection of the most appropriate set of independent variables. Including too many predictors may lead to overfitting, while including too few may omit important explanatory factors. Therefore, model selection techniques are used to identify a

suitable subset of variables. Therefore, model selection helps in building a regression model that is both **efficient** and **interpretable**.

The three common model selection techniques are **forward selection**, **backward elimination**, and **stepwise selection**.

- **Forward selection** begins with no independent variables in the model and then adds variables one by one based on their contribution to improving model performance.
- **Backward elimination** starts with all candidate variables included and removes those that do not contribute significantly to the model.
- **Stepwise selection** combines both forward and backward procedures, adding and removing variables iteratively in order to arrive at a balanced and efficient model.

These techniques are important because they help in developing a regression model that is both interpretable and predictive.

These can be understood clearly with a small numerical example.

Example 3: Suppose a company wants to predict **monthly sales (Y)** using three possible predictors:

- X_1 : Advertising spend
- X_2 : Number of sales calls
- X_3 : Average temperature

The observed data are:

Observation	Advertising X_1	Sales Calls X_2	Temperature X_3	Sales Y
1	2	5	30	40
2	4	6	32	50
3	6	7	31	65
4	8	8	29	78
5	10	9	30	90

Now we need to prepare a table to tabulate the **Candidate Models and Their Results**. That table is usually **prepared after fitting several regression models separately** and then comparing their performance measures. So, Let's understand **How the “Candidate Models and Their Results” table is prepared**

Solution: Suppose we have:

- Y = Sales
- X_1 = Advertising
- X_2 = Sales Calls
- X_3 = Temperature

To prepare the table, we do **not** start with one final model directly. Instead, we create **different possible regression models** using different combinations of predictors, and then compute measures like R^2 , **Adjusted R^2** , and **p-values**, sometimes **AIC**, **BIC**, **MSE**, **RMSE**, etc. also.

Step 1: List all possible candidate models: With 3 predictors (X_1, X_2, X_3), the possible models are:

Model	Predictors Included
M0	None (only intercept)
M1	X_1
M2	X_2
M3	X_3
M4	X_1, X_2
M5	X_1, X_3
M6	X_2, X_3
M7	X_1, X_2, X_3

This is exactly why the candidate model table had 8 rows.

Step 2: Fit each model separately: Now each model is estimated using regression.

For example:

Model M1: $Y = \beta_0 + \beta_1 X_1$

Model M2: $Y = \beta_0 + \beta_2 X_2$

Model M4: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Model M7: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Each time, software or manual regression calculations are used to estimate the coefficients.

Step 3: Compute model evaluation measures: After fitting each model, we compute:

1. **R^2 :** It tells us how much variation in Y is explained by that model.

$$R^2 = 1 - \frac{SSE}{SST}$$

where:

- SSE = Sum of Squared Errors
- SST = Total Sum of Squares

2. **Adjusted R^2 :** This adjusts R^2 for the number of predictors, because plain R^2 usually increases when more variables are added.

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

where:

- n = number of observations
- k = number of predictors

This is why Adjusted R^2 is often more useful for model selection.

3. **p-values:** These are computed for each coefficient to check whether a predictor significantly contributes to the model.

Step 4: Put the results into one comparison table

After estimating each candidate model, the results are summarized into one table like this:

Model	Predictors Included	R^2	Adjusted R^2	Remark
-------	---------------------	-------	----------------	--------

Model	Predictors Included	R^2	Adjusted R^2	Remark
M0	None (intercept only)	0.000	0.000	Baseline model
M1	X_1	0.978	0.971	Very strong predictor
M2	X_2	0.930	0.907	Good predictor
M3	X_3	0.050	-0.267	Very weak predictor
M4	X_1, X_2	0.985	0.970	Slight improvement
M5	X_1, X_3	0.979	0.958	No real gain from X_3
M6	X_2, X_3	0.932	0.864	X_3 adds little
M7	X_1, X_2, X_3	0.986	0.944	Model becomes more complex

So that table is basically a **summary table of all separately fitted models**.

Let's Understand How is one row actually prepared? :Take Model M1 as an example.

Suppose the data are:

Obs	X_1	Y
1	2	40
2	4	50
3	6	65
4	8	78
5	10	90

We fit: $Y = \beta_0 + \beta_1 X_1$;

After calculation, suppose we get: $\hat{Y} = 27.6 + 6.3X_1$

Then we calculate:

- predicted values $\hat{Y}$
- residuals $(Y - \hat{Y})$
- $SSE = \sum (Y - \hat{Y})^2$
- $SST = \sum (Y - \bar{Y})^2$
- $R^2 = 1 - \frac{SSE}{SST}$
- Adjusted R^2

Then the row for M1 is filled.

So, the row is not guessed. It is based on actual regression calculations.

Now suppose, after fitting regression models in software, the following results are obtained.

Model	Predictors Included	R^2	Adjusted R^2	Remark
M0	None (intercept only)	0.000	0.000	Baseline model
M1	X_1	0.978	0.971	Very strong predictor
M2	X_2	0.930	0.907	Good predictor
M3	X_3	0.050	-0.267	Very weak predictor
M4	X_1, X_2	0.985	0.970	Slight improvement
M5	X_1, X_3	0.979	0.958	No real gain from X_3
M6	X_2, X_3	0.932	0.864	X_3 adds little
M7	X_1, X_2, X_3	0.986	0.944	Model becomes more complex

Also suppose the p-values in the full model $M7$ are:

- $X_1: p = 0.01$
- $X_2: p = 0.08$
- $X_3: p = 0.72$

These values suggest that X_1 is highly important, X_2 is somewhat useful, and X_3 is not useful.

1. Forward Selection: In forward selection, we begin with **no predictor** in the model and add variables one at a time.

Step 1: Start with no variable: Initial model $Y = \beta_0$; This is only the intercept model.

Step 2: Add the best single predictor: We compare one-variable models:

- X_1 : Adjusted $R^2 = 0.971$
- X_2 : Adjusted $R^2 = 0.907$
- X_3 : Adjusted $R^2 = -0.267$

Since X_1 gives the best performance, it is added first.

Selected model: $Y = \beta_0 + \beta_1 X_1$

Step 3: Test remaining variables: Now we check whether adding X_2 or X_3 improves the model.

- Add X_2 : Adjusted $R^2 = 0.970$
- Add X_3 : Adjusted $R^2 = 0.958$

Neither gives a meaningful improvement over 0.971. In fact, the adjusted R^2 does not improve. So, the **Final Forward Selection Model** is $Y = \beta_0 + \beta_1 X_1$

It can be interpreted that, Forward selection chooses Advertising spend (X_1) as the only predictor because it explains sales extremely well by itself. Adding the other variables does not improve the model enough to justify extra complexity.

2. Backward Elimination: In backward elimination, we start with **all predictors** and remove them one by one.

Step 1: Start with full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

From the full model:

- $X_1: p = 0.01$
- $X_2: p = 0.08$
- $X_3: p = 0.72$

Since X_3 has the highest p-value and is clearly not significant, we remove X_3 .

Step 2: Fit reduced model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Now suppose in this reduced model:

- $X_1: p = 0.01$
- $X_2: p = 0.06$

X_2 is borderline, but the model remains acceptable and interpretable.

So, the Final Backward Elimination Model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

It can be interpreted that, Backward elimination removes Temperature (X_3) because it does not contribute to predicting sales. The final model keeps Advertising and Sales Calls, suggesting that both business-related variables are useful, while weather is not relevant here.

3. Stepwise Selection: Stepwise selection is a combination of forward selection and backward elimination. A variable can be added if it helps the model, and later removed if it becomes unnecessary.

Step 1: Start with no predictor: As before, X_1 is added first because it is the strongest individual predictor.

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

Step 2: Consider adding another variable: Among the remaining variables, X_2 provides some additional improvement, so it is added.

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

Step 3: Recheck all included variables: After adding X_2 , both X_1 and X_2 are checked again. Both remain reasonably useful, so both are kept.

Step 4: Test X_3 : Adding X_3 does not improve the model and its p-value is very high, so it is not included.

Final Stepwise Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

It can be interpreted that, Stepwise selection arrives at a balanced model using Advertising and Sales Calls. It excludes Temperature because it does not make a meaningful contribution.

Comparison of Results

<u>Technique</u>	<u>Final Selected Model</u>
Forward Selection	$Y = \beta_0 + \beta_1 X_1$
Backward Elimination	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
Stepwise Selection	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Overall Interpretation of the Results produced in this example shows that Advertising spend(X_1) is the most important predictor of sales. It alone explains a very large proportion of variation in sales. Sales calls (X_2) add some extra value, though their contribution is smaller. Temperature (X_3) does not significantly help in explaining sales and is therefore excluded in most cases.

The example also shows why model selection techniques are useful:

- Forward selection tends to produce a simpler model.
- Backward elimination starts with everything and removes unimportant variables.
- Stepwise selection balances both addition and removal of predictors.

In practice, these techniques help researchers and analysts choose a model that is not only statistically sound but also easier to interpret and use for prediction.

The discussion on model selection concludes with the understanding that model selection techniques are essential in multiple regression because they help identify the most suitable subset of predictors. In the present example, all three methods confirmed that Temperature was not useful, while Advertising was the strongest predictor. Depending on the technique used, the final model may be simpler or slightly more comprehensive, but the main objective remains the same: to obtain a regression model that is both predictive and interpretable.

Finally, it can be understood from the above discussion that Multiple regression has wide applications in both business and scientific research because many practical problems depend on several explanatory factors at the same time. In marketing, it is used to study the effect of price, promotion, and placement on sales performance. In healthcare, it helps examine how age, treatment type, and comorbidities influence patient outcomes. In finance, multiple regression supports credit scoring models by considering factors such as income, debt-to-income ratio, and credit history. In human resources, it is used for salary determination based on experience, education, and performance indicators. These examples show that multiple

regression is a versatile and powerful method for both explanation and prediction in complex real-world situations.

This model is widely used in predictive analytics because real-world outcomes are rarely influenced by a single factor. For instance, house price prediction may depend on area, location, number of bedrooms, age of the property, and nearby amenities. Multiple linear regression allows all these predictors to be considered simultaneously.

The key benefit of multiple regression is that it provides a more realistic and comprehensive explanation of the dependent variable. It also allows the analyst to measure the individual contribution of each predictor while controlling for the others. However, it requires careful attention to assumptions such as linearity, independence of errors, homoscedasticity, normality of residuals, and absence of multicollinearity. If these assumptions are violated, the accuracy and reliability of the model may be affected.

☛ Check Your Progress 3

- Q1. What is Multiple Linear Regression? Explain its importance.
.....
.....
- Q2. Write the general form of a Multiple Linear Regression model and explain its components.
.....
.....
- Q3. Explain the interpretation of regression coefficients in Multiple Regression.
.....
.....
- Q4. What is Multicollinearity? How is it detected?
.....
.....
- Q5. Explain model selection techniques used in Multiple Regression.
.....
.....

8.3.3POLYNOMIAL REGRESSION

Polynomial regression is an important extension of linear regression that is used when the relationship between the independent variable and the dependent variable is not adequately represented by a straight line. In many real-world datasets, the pattern of association between variables is curved rather than linear. In such situations, simple linear regression may fail to provide a satisfactory fit. Polynomial regression addresses this limitation by incorporating higher-order powers of the independent variable into the regression equation, thereby allowing the model to capture non-linear relationships more effectively. Although the relationship between the variables becomes curved, polynomial regression is still considered a type of regression analysis because it remains linear in terms of the coefficients.

The general form of a polynomial regression model of degree d is expressed as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \dots + \beta_d X^d + \epsilon$$

In this equation, Y represents the dependent variable, X is the independent variable, $\beta_0, \beta_1, \beta_2, \dots, \beta_d$ are the regression coefficients, and ϵ is the error term. The inclusion of squared, cubed, and higher-order terms allows the regression curve to bend and adapt to the underlying structure of the data. A second-degree polynomial, also called a quadratic model, takes the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

A third-degree polynomial, or cubic model, is written as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

As the degree of the polynomial increases, the model becomes more flexible and capable of fitting increasingly complex curved patterns. However, greater flexibility must be balanced against the risk of overfitting.

A crucial question is, When to Use Polynomial Regression? It is to be noted that Polynomial regression is appropriate when the relationship between the variables clearly exhibits curvature and cannot be explained well by a simple straight-line model. One indication for its use arises when residual plots from linear regression show systematic patterns rather than random scatter, suggesting that the linear model is missing important non-linear structure. Polynomial regression is also useful when prior domain knowledge suggests that the underlying process is non-linear in nature. In addition, exploratory data analysis, particularly scatter plots, may reveal curved trends in the data that make polynomial modeling a more suitable choice. Thus, polynomial regression is selected when a linear model is too restrictive, and a curved function is needed to represent the relationship more accurately.

The graphical representation of polynomial regression typically begins with a scatter plot of data points, just as in linear regression. However, in this case the observations do not cluster around a straight-line trend; instead, they follow a curved pattern. The fitted model is represented by a smooth curved line, which serves as the polynomial line of best fit. This curve is generated by the polynomial equation and is able to reflect non-linear movement in the dependent variable as the independent variable changes. The degree of the polynomial determines the number of bends that the curve can display. For instance, a quadratic polynomial can capture one curve, while a cubic polynomial can model more complex changes with up to two bends. As a result, polynomial regression provides a more realistic representation of many naturally occurring phenomena than linear regression.

To understand **polynomial regression** clearly, let us consider a numerical example in which the relationship between the independent variable and dependent variable is **curved** rather than linear.

Example 4: Suppose a researcher wants to study the relationship between: X : Hours of practice and Y : Performance score. The observed data are:

Observation Hours of Practice X Performance Score Y

1	1	52
2	2	58
3	3	68
4	4	82
5	5	100

From the pattern of data, it appears that the increase in performance is not constant. The score is increasing at an increasing rate, which suggests a curved relationship. Therefore, instead of simple linear regression, we use a quadratic polynomial regression model.

Solution: The Polynomial Regression Model of degree 2 is written as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

where:

- Y = dependent variable
- X = independent variable
- β_0 = intercept
- β_1 = coefficient of X
- β_2 = coefficient of X^2
- ϵ = error term

For this example, the fitted model will be:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

Step 1: Prepare the Calculation Table:

Since this is a quadratic model, we need the values of X^2 .

Observation	X	Y	X^2
1	1	52	1
2	2	58	4
3	3	68	9
4	4	82	16
5	5	100	25

Now compute the required totals:

$$\sum X = 1 + 2 + 3 + 4 + 5 = 15$$

$$\sum Y = 52 + 58 + 68 + 82 + 100 = 360$$

$$\sum X^2 = 1 + 4 + 9 + 16 + 25 = 55$$

$$\sum X^3 = 1 + 8 + 27 + 64 + 125 = 225$$

$$\sum X^4 = 1 + 16 + 81 + 256 + 625 = 979$$

$$\sum XY = 52 + 116 + 204 + 328 + 500 = 1200$$

$$\sum X^2Y = 52 + 232 + 612 + 1312 + 2500 = 4708$$

Step 2: Form the Normal Equations

For quadratic polynomial regression, the normal equations are:

$$\sum Y = n\beta_0 + \beta_1\sum X + \beta_2\sum X^2$$

$$\sum XY = \beta_0\sum X + \beta_1\sum X^2 + \beta_2\sum X^3$$

$$\sum X^2Y = \beta_0\sum X^2 + \beta_1\sum X^3 + \beta_2\sum X^4$$

Substituting the values:

$$360 = 5\beta_0 + 15\beta_1 + 55\beta_2$$

$$1200 = 15\beta_0 + 55\beta_1 + 225\beta_2$$

$$4708 = 55\beta_0 + 225\beta_1 + 979\beta_2$$

Step 3: Solve the Equations

The system of equations is:

$$5\beta_0 + 15\beta_1 + 55\beta_2 = 360$$

$$15\beta_0 + 55\beta_1 + 225\beta_2 = 1200$$

$$55\beta_0 + 225\beta_1 + 979\beta_2 = 4708$$

Solving these equations gives:

$$\beta_0 = 50$$

$$\beta_1 = 1$$

$$\beta_2 = 2$$

Step 4: Write the Fitted Polynomial Regression Equation

Substituting the estimated coefficients into the model: $\hat{Y} = 50 + 1X + 2X^2$

So, the fitted quadratic polynomial regression equation is: $\hat{Y} = 50 + X + 2X^2$

Step 5: Compute the Predicted Values

Now use the fitted equation to calculate the predicted values.

$$\text{For } X = 1: \hat{Y} = 50 + 1(1) + 2(1^2) = 50 + 1 + 2 = 53$$

$$\text{For } X = 2: \hat{Y} = 50 + 1(2) + 2(2^2) = 50 + 2 + 8 = 60$$

$$\text{For } X = 3: \hat{Y} = 50 + 1(3) + 2(3^2) = 50 + 3 + 18 = 71$$

$$\text{For } X = 4: \hat{Y} = 50 + 1(4) + 2(4^2) = 50 + 4 + 32 = 86$$

$$\text{For } X = 5: \hat{Y} = 50 + 1(5) + 2(5^2) = 50 + 5 + 50 = 105$$

So the predicted values are:

X	Actual Y	Predicted $\hat{Y}$
1	52	53
2	58	60
3	68	71

X	Actual Y	Predicted $\hat{Y}$
4	82	86
5	100	105

Step 6: Calculate Residuals: Residual $Y - \hat{Y}$

X	Actual Y	Predicted $\hat{Y}$	Residual $Y - \hat{Y}$
1	52	53	-1
2	58	60	-2
3	68	71	-3
4	82	86	-4
5	100	105	-5

These residuals show the deviation of observed values from the fitted polynomial curve.

Step 7: Interpretation of the Polynomial Regression Equation: The fitted polynomial model is: $\hat{Y} = 50 + X + 2X^2$

- Interpretation of the Intercept ($\beta_0 = 50$): The intercept indicates the estimated value of performance score when $X = 0$. Thus, when practice hours are zero, the predicted score is 50.
- Interpretation of the Linear Coefficient ($\beta_1 = 1$): The coefficient of X shows the linear part of the relationship. It contributes positively to the score, meaning that as practice hours increase, the performance score also tends to increase.
- Interpretation of the Quadratic Coefficient ($\beta_2 = 2$): The coefficient of X^2 is positive, which means the curve bends upward. This indicates that the increase in performance score becomes larger as practice hours increase. In other words, the effect of practice is not constant; it accelerates with more hours of practice.

This is the key reason for using polynomial regression here instead of simple linear regression.

Step 8: Interpretation of Results: The fitted model suggests that there is a **non-linear positive relationship** between practice hours and performance score. Initially, the score increases slowly, but as practice hours increase further, the score rises more rapidly. This curved trend is captured by the positive quadratic term $2X^2$.

Thus, the model indicates that:

- performance improves as practice increases,

- the rate of improvement is not constant,
- the improvement becomes stronger at higher levels of practice.

This makes polynomial regression more suitable than linear regression for such data.

Step 9: Prediction Using the Model: Suppose we want to predict the performance score for a student who practices for **6 hours**. Using the fitted equation:

$$\hat{Y} = 50 + 1(6) + 2(6^2)$$

$$\hat{Y} = 50 + 6 + 72 = 128$$

So, the predicted performance score for **6 hours of practice** is:

$$\hat{Y} = 128$$

This numerical example demonstrates how **polynomial regression** is used when the relationship between variables is curved rather than linear. Using the quadratic model,

$$\hat{Y} = 50 + X + 2X^2$$

we were able to model the non-linear relationship between hours of practice and performance score. The positive quadratic term shows that performance increases at an increasing rate as practice hours rise.

Therefore, polynomial regression is especially useful when:

- a straight line does not fit the data well,
- the scatter plot shows curvature,
- the rate of change in the dependent variable is not constant.

This makes polynomial regression an important tool in predictive data analysis for modeling real-world non-linear relationships.

Through this numerical example we learned about the polynomial regression and interpreted the results, now we need to learn about the **Advantages and Applications of Polynomial regression**, it offers several important advantages, a few are listed below:

- One major advantage is that it is capable of capturing non-linear relationships, making it more flexible than simple linear regression. At the same time, it remains relatively simple to implement, especially with modern statistical software.
- Polynomial regression can model a wide range of curve shapes, and when low-degree polynomials are used, the model often remains reasonably interpretable. This balance between flexibility and simplicity makes it a valuable technique in predictive data analysis.

Polynomial regression is applied in many practical fields:

- In **biology**, it is used to model the growth rates of organisms over time, where growth may accelerate and then slow down.

- In **economics**, it helps analyze diminishing returns in production functions, where increases in input do not always lead to proportionate increases in output.
- In **engineering**, polynomial regression is useful for modeling stress-strain relationships in materials, which are often non-linear.
- In **environmental science**, it can be applied to study changes in pollutant concentration over time.
- In **finance**, it is used for modeling situations such as option pricing, where payoffs may follow non-linear structures.

These applications demonstrate that polynomial regression is especially suitable in contexts where the effect of the predictor changes at different levels.

Polynomial regression is especially useful in situations where the process under study changes direction or intensity over time. Like in the following cases:

- In **disease progression modeling**, polynomial regression helps represent epidemic curves by capturing the initial growth phase, the peak level, and the eventual decline of disease outbreaks.
- In **tissue growth analysis**, particularly in medical research, polynomial models are employed to study non-linear growth patterns, such as tumor development or tissue regeneration.
- In **automotive engineering**, polynomial regression can be used in speed control systems of autonomous vehicles to model non-linear relationships between environmental conditions and appropriate speed adjustments.

These examples highlight the practical importance of polynomial regression in fields where simple linear assumptions are too restrictive.

Despite its strengths, *polynomial regression presents several challenges that must be considered carefully. Some of the major challenges are:*

- Challenge of **overfitting**. Higher-degree polynomials may fit the training data extremely well, but such models may fail to generalize to new data and may produce poor predictive performance. This risk can be managed by using cross-validation to select an appropriate polynomial degree, applying regularization methods, and keeping the degree as low as possible while still capturing the essential shape of the data.
- Another challenge is **extrapolation**. Polynomial models may behave unpredictably outside the range of the observed data, often generating unrealistic predictions at extreme values of the independent variable. Therefore, caution must be exercised when using polynomial regression for forecasting beyond the available data range.
- A further limitation concerns **interpretability**. While low-degree polynomial models may still be understandable, higher-degree polynomials become increasingly difficult to interpret in a meaningful way. This reduces one of the main advantages traditionally associated with regression analysis, namely the ability to explain

relationships clearly. Hence, polynomial regression should be applied with care, balancing flexibility, predictive accuracy, and interpretability.

☛ Check Your Progress 4

Q1. What is Polynomial Regression? Why is it used?.

.....

Q2. Write the general form of a Polynomial Regression model.

.....

Q3. How does Polynomial Regression differ from Linear Regression?

.....

Q4. What are the advantages and limitations of Polynomial Regression?

.....

Q5. When should Polynomial Regression be preferred over Linear Regression?

.....

8.3.4 RIDGE AND LASSO REGRESSION

Ridge regression and Lasso regression are regularization techniques used to improve the performance of regression models, especially when there are many independent variables or when multicollinearity exists among the predictors. In ordinary linear regression, highly correlated predictors can make coefficient estimates unstable and may lead to overfitting. Ridge and Lasso address this issue by adding a penalty term to the regression model, which shrinks the coefficient values and helps produce a more reliable model.

The main difference between the two methods is that ridge regression shrinks coefficients but keeps all variables, whereas lasso regression can shrink some coefficients to zero and thus simplify the model. Ridge is preferred when most predictors are relevant but highly correlated, while lasso is more suitable when only a few predictors are truly important and feature selection is needed.

Overall, both ridge and lasso regression are valuable techniques in predictive data analysis because they help reduce overfitting, improve model generalization, and handle multicollinearity effectively.

Now, in this section we will discuss both type of regressions in a bit detail, let's begin with the Ridge regression.

Ridge Regression:

Ridge regression is an advanced form of linear regression that is used when the independent variables are highly correlated with each other or when the model has a large number of predictors. In ordinary least squares regression, the coefficients are estimated in such a way that the sum of squared errors between actual and predicted values is minimized. However, when multicollinearity exists, that is, when the predictor variables are strongly related to one

another, the estimated coefficients may become unstable and may vary greatly with small changes in the data. Ridge regression addresses this problem by introducing a penalty term into the regression objective function. This penalty discourages very large coefficient values and therefore helps produce a more stable and reliable model.

The objective function of ridge regression is written as:

$$\text{Minimize } \sum (Y_i - \hat{Y}_i)^2 + \lambda \sum \beta_j^2$$

In this expression, the first term, $\sum (Y_i - \hat{Y}_i)^2$, is the usual residual sum of squares from ordinary linear regression. The second term, $\lambda \sum \beta_j^2$, is called the ridge penalty or regularization term. Here, λ is known as the **regularization parameter**, and it controls the strength of the penalty. When $\lambda = 0$, ridge regression becomes identical to ordinary least squares regression. As the value of λ increases, the penalty becomes stronger, and the regression coefficients are forced to shrink toward zero. This shrinkage reduces the effect of multicollinearity and helps avoid overfitting. However, unlike lasso regression, ridge regression does not make coefficients exactly zero, which means that all predictor variables remain in the model.

Ridge regression is especially useful when there are several predictors that carry overlapping information. In such cases, instead of trying to assign large and unstable coefficients to correlated variables, ridge regression distributes their influence more smoothly by shrinking the coefficients. This often improves the predictive performance of the model, especially on new data.

For a beginner, ridge regression can be understood in a very simple way. Ordinary linear regression tries to fit the data as closely as possible, but sometimes this results in very large or unstable coefficients, especially when predictors are strongly related to each other. Ridge regression adds a “penalty” for having large coefficients. Because of this penalty, the model prefers smaller and more balanced coefficient values. This makes the model more stable and often better at prediction.

One may think of ridge regression as telling the model: “Do not only fit the data well, but also keep the coefficient values under control.”

Let’s understand the terms associated with Ridge regression by solving a beginner-level Simple Numerical Example of Ridge Regression

Example5: Let us take. Suppose a teacher wants to predict the final marks of students on the basis of two highly related predictors:

- X_1 = Hours studied
- X_2 = Number of practice tests attempted
- Y = Final marks

Assume the following small dataset:

Student Hours Studied (X_1) Practice Tests (X_2) Final Marks (Y)

Student Hours Studied (X_1) Practice Tests (X_2) Final Marks (Y)

1	1	1	52
2	2	2	57
3	3	3	63
4	4	4	68

Here, the two predictors are perfectly related because students who study more hours also attempt more practice tests. In such a case, ordinary regression faces difficulty in deciding how much weight should be given separately to X_1 and X_2 . Ridge regression helps solve this issue by shrinking the coefficients.

For simplicity, suppose the ordinary least squares regression gives the following model:

$$\hat{Y} = 47 + 3X_1 + 2X_2$$

Now let us apply ridge regression with a penalty parameter $\lambda = 1$. For beginner-level understanding, suppose the effect of ridge penalty shrinks the coefficients from 3 and 2 to 2.5 and 1.5, respectively. Then the ridge regression model becomes:

$$\hat{Y}_{ridge} = 47 + 2.5X_1 + 1.5X_2$$

Thus, the coefficients are reduced, but they are not removed from the model.

Now, let's perform Stepwise Prediction Using the Ridge Model. Let us calculate the predicted marks for each student using the ridge regression equation.

For Student 1

$$\begin{aligned}\hat{Y} &= 47 + 2.5(1) + 1.5(1) \\ \hat{Y} &= 47 + 2.5 + 1.5 = 51\end{aligned}$$

For Student 2

$$\begin{aligned}\hat{Y} &= 47 + 2.5(2) + 1.5(2) \\ \hat{Y} &= 47 + 5 + 3 = 55\end{aligned}$$

For Student 3

$$\begin{aligned}\hat{Y} &= 47 + 2.5(3) + 1.5(3) \\ \hat{Y} &= 47 + 7.5 + 4.5 = 59\end{aligned}$$

For Student 4

$$\begin{aligned}\hat{Y} &= 47 + 2.5(4) + 1.5(4) \\ \hat{Y} &= 47 + 10 + 6 = 63\end{aligned}$$

So the predicted values are:

Student	Actual Marks (Y)	Predicted Marks ($\hat{Y}_{ridge}$)
1	52	51

Student	Actual Marks (Y)	Predicted Marks ($\hat{Y}_{ridge}$)
2	57	55
3	63	59
4	68	63

The ridge regression model is: $\hat{Y} = 47 + 2.5X_1 + 1.5X_2$

This equation suggests that both hours studied and practice tests positively affect final marks. Holding the number of practice tests constant, a one-unit increase in hours studied increases the predicted marks by 2.5 marks. Similarly, holding hours studied constant, a one-unit increase in practice tests increases the predicted marks by 1.5 marks. The intercept value of 47 indicates the predicted marks when both predictors are zero.

The main point to observe here is that the coefficients in ridge regression are smaller than those in the ordinary regression model. This shrinkage is intentional. Because the predictors are highly correlated, ordinary regression may produce unstable estimates, whereas ridge regression stabilizes them by reducing their size. In this way, ridge regression reduces the risk of overfitting and improves the reliability of the model.

Thus, it can be concluded from the above numerical that Ridge regression is a useful extension of linear regression that helps deal with multicollinearity and improves the stability of the model. It works by adding a penalty term based on the squared values of the coefficients. As the regularization parameter increases, the coefficients are shrunk more strongly toward zero, but they never become exactly zero. Therefore, all variables remain part of the model. In the above numerical example, ridge regression reduced the sizes of the coefficients for hours studied and practice tests, thereby producing a more balanced model. For students, the key takeaway is that ridge regression is especially helpful when predictor variables are highly related and ordinary regression becomes unreliable.

Lasso Regression

Lasso regression, which stands for **Least Absolute Shrinkage and Selection Operator (Lasso)**, is a regularized regression technique used to improve the performance of prediction models, especially when the dataset contains a large number of independent variables. In ordinary linear regression, the coefficients are estimated only by minimizing the sum of squared errors between the actual values and the predicted values. However, when the number of predictors is large, or when several predictors are weak, irrelevant, or highly correlated, ordinary regression may produce unstable estimates and may overfit the training data. Lasso regression addresses this issue by introducing a penalty term into the objective function, thereby controlling the size of the regression coefficients and simplifying the model.

The objective function of lasso regression is:

$$\text{Minimize } \sum (Y_i - \hat{Y}_i)^2 + \lambda \sum |\beta_j|$$

In this expression, the first term, $\sum (Y_i - \hat{Y}_i)^2$, represents the residual sum of squares, which measures how closely the predicted values match the actual observations. The second term, $\lambda \sum |\beta_j|$, is the **lasso penalty**, where λ is the regularization parameter and $|\beta_j|$ denotes the absolute value of the regression coefficients. The parameter λ controls the strength of the penalty. When $\lambda = 0$, lasso regression becomes the same as ordinary least squares regression. As the value of λ increases, the penalty becomes stronger, and the regression coefficients are shrunk more aggressively.

In simple terms, lasso regression can be understood as a method that tells the model: “Fit the data well, but also remove variables that are not really helping.”

The most distinctive feature of lasso regression is that it can shrink some coefficients **exactly to zero**. This happens because the penalty is based on the absolute values of the coefficients rather than their squares. As a result, lasso not only reduces the size of coefficients but can also completely remove some predictors from the model. In this way, lasso performs **automatic variable selection**. This makes the model simpler, easier to interpret, and often more effective when dealing with datasets that contain many variables but only a few truly important ones.

This property distinguishes lasso regression from ridge regression. Ridge regression shrinks coefficients toward zero but usually keeps all variables in the model, whereas lasso regression may shrink some coefficients completely to zero and exclude the corresponding variables. Therefore, lasso is especially useful when the goal is not only prediction but also identification of the most relevant predictors.

Lasso regression is widely applied in situations where a dataset has a large number of potential explanatory variables. For example, in marketing analytics, a business may have dozens of customer-related features such as age, income, purchase history, browsing behaviour, discount usage, and demographic characteristics. Not all these variables are equally important for predicting customer spending. Lasso regression can identify the variables that truly contribute to the prediction while removing those that add little value. This makes it highly useful in machine learning, business analytics, bioinformatics, finance, and other areas involving high-dimensional data.

Because lasso regression reduces overfitting and improves interpretability, it is particularly valuable in predictive data analysis. By shrinking unimportant coefficients to zero, it creates a more compact model that is easier to understand and often better at generalizing to new data. For this reason, lasso regression is considered one of the most important regularization techniques in modern statistical learning and applied machine learning.

Let’s understand the terms associated with Lasso regression by solving a beginner-level Simple Numerical Example of Lasso Regression

A Simple Numerical Illustration is as follows:

Suppose a researcher wants to predict student performance using the following three predictors:

- X_1 : Hours studied
- X_2 : Attendance
- X_3 : Number of extracurricular activities

Assume that after fitting a regression model, the estimated coefficients before regularization are:

$$\hat{Y} = 40 + 5X_1 + 3X_2 + 0.5X_3$$

Here, the coefficient of extracurricular activities is very small compared with the other two predictors, suggesting that it may not be very important.

Now suppose lasso regression is applied with a suitable value of λ . After penalization, the model becomes:

$$\hat{Y}_{lasso} = 40 + 4.2X_1 + 2.6X_2 + 0X_3$$

This means that lasso has shrunk the coefficient of X_3 exactly to zero. As a result, the variable X_3 is automatically removed from the model, and the final simplified regression equation becomes:

$$\hat{Y}_{lasso} = 40 + 4.2X_1 + 2.6X_2$$

This illustrates the main advantage of lasso regression: it retains important variables while excluding unimportant ones. This combination of coefficient shrinkage and variable selection is what makes lasso regression especially powerful in modern predictive modelling.

The final lasso model indicates that **hours studied** and **attendance** are important predictors of student performance, while **extracurricular activities** do not contribute significantly in the presence of the other variables. By shrinking the third coefficient to zero, lasso produces a simpler and more interpretable model. This not only helps in reducing unnecessary complexity but also improves understanding of which variables truly matter in the prediction task.

One more Simple Numerical Example of Lasso Regression: Let us consider one simpler example suitable for beginner-level students.

Suppose a teacher wants to predict the **final marks** of students using the following three predictors:

- X_1 = Hours studied
- X_2 = Attendance score
- X_3 = Number of extracurricular activities

The dependent variable is:

- Y = Final marks

Assume that, using ordinary linear regression, the following model is obtained:

$$\hat{Y} = 30 + 4X_1 + 3X_2 + 0.5X_3$$

This equation means:

- each extra hour of study increases marks by 4,
- each one-unit increase in attendance increases marks by 3,

- each one-unit increase in extracurricular activities increases marks by only 0.5.

From these coefficients, it already appears that X_3 has a very small contribution compared to the other variables.

Now let us apply **lasso regression** with a penalty parameter $\lambda = 1$. For beginner-level understanding, suppose that after applying the lasso penalty, the coefficients change as follows:

- coefficient of X_1 : from 4 to 3.5
- coefficient of X_2 : from 3 to 2.5
- coefficient of X_3 : from 0.5 to 0

So, the lasso regression model becomes:

$$\hat{Y}_{lasso} = 30 + 3.5X_1 + 2.5X_2 + 0X_3$$

or simply,

$$\hat{Y}_{lasso} = 30 + 3.5X_1 + 2.5X_2$$

This shows that lasso has removed the extracurricular activities variable from the model by shrinking its coefficient exactly to zero.

Let's understand the Stepwise procedure of Prediction Using the Lasso Model, with the help of an example: Suppose we have the following student data:

Student	Hours Studied (X_1)	Attendance (X_2)	Extracurricular Activities (X_3)	Actual Marks (Y)
1	2	3	5	45
2	4	4	2	54
3	5	5	1	60
4	6	6	3	66

Using the lasso regression model:

$$\hat{Y}_{lasso} = 30 + 3.5X_1 + 2.5X_2$$

we calculate the predicted values.

For Student 1

$$\hat{Y} = 30 + 3.5(2) + 2.5(3)$$

$$\hat{Y} = 30 + 7 + 7.5 = 44.5$$

For Student 2

$$\hat{Y} = 30 + 3.5(4) + 2.5(4)$$

$$\hat{Y} = 30 + 14 + 10 = 54$$

For Student 3

$$\hat{Y} = 30 + 3.5(5) + 2.5(5)$$

$$\hat{Y} = 30 + 17.5 + 12.5 = 60$$

For Student 4

$$\hat{Y} = 30 + 3.5(6) + 2.5(6)$$

$$\hat{Y} = 30 + 21 + 15 = 66$$

Thus, the predicted values are:

Student	Actual Marks (Y)	Predicted Marks ($\hat{Y}_{lasso}$)
1	45	44.5
2	54	54
3	60	60
4	66	66

Now, we need to Understand that How the Lasso Penalty Works?

Let us now see how the penalty is applied in a simple way.

From the ordinary regression model:

$$\hat{Y} = 30 + 4X_1 + 3X_2 + 0.5X_3$$

the lasso penalty part is:

$$\lambda \sum |\beta_j|$$

Since $\lambda = 1$,

$$\text{Penalty} = |4| + |3| + |0.5| = 4 + 3 + 0.5 = 7.5$$

Now suppose the lasso model becomes:

$$\hat{Y}_{lasso} = 30 + 3.5X_1 + 2.5X_2 + 0X_3$$

Then the new penalty is:

$$|3.5| + |2.5| + |0| = 3.5 + 2.5 + 0 = 6$$

So the penalty has been reduced from **7.5 to 6**.

This is the key idea of lasso regression. It tries to balance two things at the same time:

1. fit the data well
2. keep the coefficients small

If a variable contributes very little, lasso may reduce its coefficient completely to zero in order to reduce the penalty and simplify the model.

The Interpretation of the Results from the lasso regression model i.e.:

$$\hat{Y}_{lasso} = 30 + 3.5X_1 + 2.5X_2$$

This equation shows that:

- for every additional hour of study, marks increase by **3.5 marks**, keeping attendance constant,
- for every one-unit increase in attendance, marks increase by **2.5 marks**, keeping hours studied constant,
- extracurricular activities do not make a meaningful contribution in the presence of the other variables, so their coefficient becomes zero.

This means that **hours studied** and **attendance** are the truly important predictors of final marks in this example, while **extracurricular activities** do not significantly help in prediction.

It can be understood that, lasso regression can be understood very simply, by understanding the following:

- Ordinary regression says: “Use all variables to fit the data as well as possible.”
- Lasso regression says: “Fit the data well, but also remove variables that are not really useful.”

So, lasso acts like a filter. It keeps the important variables and drops the less important ones. That is why lasso regression is very helpful when there are many predictors and the analyst wants a model that is both **accurate** and **easy to understand**.

In the numerical example above, lasso regression identified that **hours studied** and **attendance** were useful predictors of student marks, while **extracurricular activities** were not important enough to remain in the model. Therefore, the final model became simpler and more interpretable. For students, the main idea to remember is that lasso regression helps build a regression model that is not only predictive but also selective about which variables truly matter.

This section concludes that Lasso regression is a powerful extension of ordinary linear regression. It improves the model by adding a penalty based on the absolute values of the coefficients. This penalty shrinks coefficient values, and in some cases it reduces them exactly to zero. As a result, lasso regression not only reduces overfitting but also performs automatic variable selection.

☛ Check Your Progress 5

Q1. What are Ridge and Lasso Regression? Why are they used?

.....

Q2. Differentiate between Ridge Regression and Lasso Regression.

.....

Q3. Explain the concept of regularization in regression.

.....

Q4. How do Ridge and Lasso Regression help in handling multicollinearity?

.....

Q5. What are the advantages and limitations of Ridge and Lasso Regression?

.....

8.3.5 SUPPORT VECTOR REGRESSION

Support Vector Regression, commonly called **SVR**, is a regression technique derived from the concept of Support Vector Machines. While Support Vector Machines are widely known for classification problems, the same idea can also be applied to regression problems where

the goal is to predict a continuous numerical value. SVR is especially useful when the relationship between variables is complex and when the analyst wants a model that is not overly affected by noise in the data.

The basic idea of Support Vector Regression is different from ordinary linear regression. In linear regression, the model tries to minimize the total prediction error for all data points. In SVR, instead of trying to make every error as small as possible, the model attempts to fit a function in such a way that errors within a certain acceptable range are ignored. This acceptable range is called the **epsilon-insensitive tube**. In simple words, if the predicted value lies within a distance of ϵ above or below the actual value, then that error is treated as acceptable and no penalty is given. Only those points that fall outside this tube contribute to the error.

Thus, SVR tries to find a regression line or curve that is as flat as possible while ensuring that most observations lie within the epsilon margin. The observations that lie on the boundary of the tube or outside it are called **support vectors**, because these are the points that actually determine the position of the regression function.

The Support Vector Regression Model

For a simple linear case, the Support Vector Regression function is written as:

$$f(x) = wx + b$$

where:

- x is the independent variable,
- w is the weight or slope,
- b is the intercept.

The objective of SVR is to minimize:

$$\frac{1}{2} \|w\|^2$$

subject to the condition that the prediction errors remain within the epsilon tube as much as possible. When points fall outside the tube, slack variables are introduced, and the optimization becomes:

$$\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum (\xi_i + \xi_i^*)$$

subject to:

$$\begin{aligned} Y_i - (wx_i + b) &\leq \epsilon + \xi_i \\ (wx_i + b) - Y_i &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

Here:

- ϵ defines the width of the error tolerance tube,
- C is the penalty parameter,
- ξ_i and ξ_i^* are slack variables for points outside the tube.

It is important to understand that **SVR allows small errors but penalizes only larger errors**.

A regular question that requires attention is, Why Support Vector Regression is Useful?

The usefulness of SVR will be clearer, after understanding that the SVR balances the two goals, given below i.e.:

- keeping the model as simple and flat as possible,
- allowing small errors that do not significantly affect prediction quality.

This makes SVR effective when the data contains noise or when the relationship is not perfectly linear. It can also be extended to nonlinear regression using kernel functions, but it is best to first study the simple linear idea.

A Simple Numerical Example of Support Vector Regression is as follows:

Example6: Suppose by Using the SVR model gives the regression function:

$\hat{Y} = 5X + 45$ with prediction error $\epsilon = 2$, a teacher wants to predict **student marks** based on **hours of study**.

Let:

- X = Hours studied
- Y = Marks obtained

The observed data are:

Student	Hours Studied (X)	Marks (Y)
1	1	50
2	2	55
3	3	60
4	4	65
5	5	72

Solution: As per the given SVR model the regression function is: $\hat{Y} = 5X + 45$

This means the predicted marks are calculated as:

$$\hat{Y} = 5X + 45$$

Now given prediction error is: $\epsilon = 2$,

This means any prediction error up to **2 marks** is acceptable and will not be penalized.

Step 1: Compute Predicted Values by using the regression equation: $\hat{Y} = 5X + 45$

we get:

For $X = 1$ $\hat{Y} = 5(1) + 45 = 50$

For $X = 2$ $\hat{Y} = 5(2) + 45 = 55$

For $X = 3$ $\hat{Y} = 5(3) + 45 = 60$

For $X = 4$ $\hat{Y} = 5(4) + 45 = 65$

For $X = 5$ $\hat{Y} = 5(5) + 45 = 70$

So, the predicted values are:

X	Actual Y	Predicted $\hat{Y}$
1	50	50
2	55	55

X	Actual Y	Predicted $\hat{Y}$
3	60	60
4	65	65
5	72	70

Step 2: Compute Errors by using the Error equation: $\text{Error} = Y - \hat{Y}$

X	Actual Y	Predicted $\hat{Y}$	Error $Y - \hat{Y}$
1	50	50	0
2	55	55	0
3	60	60	0
4	65	65	0
5	72	70	2

Step 3: Apply the Epsilon Rule

In Support Vector Regression, the value of epsilon defines the acceptable margin of error around the regression line. For this example, $\epsilon = 2$.

This means that if the absolute difference between the actual value and the predicted value is **less than or equal to 2**, then that error is treated as acceptable and **no penalty** is applied.

However, if the absolute error is **greater than 2**, then the observation falls outside the epsilon-insensitive tube and a penalty is imposed.

Mathematically, the rule can be written as:

$$|Y - \hat{Y}| \leq 2 \Rightarrow \text{No penalty}$$

$$|Y - \hat{Y}| > 2 \Rightarrow \text{Penalty is applied}$$

Now let us check the error for each observation:

X	Actual Y	Predicted $\hat{Y}$	Error $Y - \hat{Y}$	Penalized?
1	50	50	0	No
2	55	55	0	No
3	60	60	0	No
4	65	65	0	No
5	72	70	2	No

Since the absolute error for every observation is less than or equal to the epsilon value of 2, all the data points lie within the epsilon-insensitive tube. Therefore, **none of the observations are penalized**. This indicates that the fitted SVR regression line is acceptable, as it keeps all prediction errors within the allowed tolerance range.

Step 4: Identify the Support Vectors: It is to be noted that in SVR, the points that lie on the epsilon boundary or outside it are the most important points. These are called **support vectors**.

Here, the point (5, 72) has error 2, which lies exactly on the epsilon boundary. So, this point acts like a support vector.

The other points lie exactly on the line, so they are also fitted well, but the point on the boundary is particularly important in defining the separation region.

Another Small Variation for Better Understanding

Suppose instead the predicted value for $X = 5$ were 68 instead of 70.

Then error would be:

$$72 - 68 = 4$$

Now:

$$|4| > 2$$

So only the excess beyond epsilon is penalized:

$$4 - 2 = 2$$

This means a penalty would apply for that point.

This shows the main feature of SVR:

- small errors inside epsilon are ignored
- only larger deviations outside epsilon matter

Interpretation of the Results are as follows: The fitted SVR model is:

$$\hat{Y} = 5X + 45$$

This equation suggests that for each additional hour of study, the predicted marks increase by **5 marks**. The intercept value **45** means that if a student studies for zero hours, the predicted marks would be 45.

In the numerical example above, the regression equation

$$\hat{Y} = 5X + 45$$

with $\epsilon = 2$ provided a good fit for the student marks data because all prediction errors remained within the acceptable range. Thus, the model successfully captured the relationship between hours studied and marks obtained.

The most important interpretation in SVR, however, is not only the slope and intercept but also the epsilon tolerance. Since $\epsilon = 2$, the model accepts prediction errors up to 2 marks without considering them serious. In this example, all prediction errors lie within this acceptable range, which means the model fits the data well according to the SVR principle.

This is different from ordinary regression, where every error is counted. SVR is more flexible because it focuses only on errors that are meaningfully large.

So, SVR can be understood in this simple way by understanding the following:

- Linear regression says: “Try to reduce every error.”
- Support Vector Regression says: “Small errors are okay; only worry about big errors.”

So, SVR creates a safe zone or tolerance band around the regression line. If a predicted value falls inside this band, the model is satisfied. Only points outside the band affect the optimization strongly.

That's why SVR is useful when the data has slight noise or variation and we do not want the model to overreact to every small difference.

Referring to the *objective of SVR that is to minimize* $\frac{1}{2} \|w\|^2 + C \sum (\xi_i + \xi_i^*)$, in this expression we need to understand the **Real Meaning of C**, another important parameter in SVR.

- A **large C** means the model gives a strong penalty to points outside the epsilon tube. So it tries harder to fit the data closely.
- A **small C** means the model is more relaxed and allows some larger deviations.

So:

- ϵ : controls how much error is acceptable
- **C** : controls how strict the model is when error becomes too large

The key idea of **SVR** to remember is that Support Vector Regression (SVR) allows small errors, ignores them, and concentrates only on large deviations, making it a useful and practical method for regression problems.

We are not going to further details of Support Vector Regression (SVR), which is a powerful regression method that predicts continuous values while allowing a certain amount of error through the epsilon-insensitive tube. Unlike ordinary regression, it does not try to minimize every small error. Instead, it focuses on finding a simple regression function that keeps most points within an acceptable error range. This makes the model more robust and less sensitive to noise.

☛ Check Your Progress 6

Q1. What is Support Vector Regression (SVR)? Explain its objective.

.....

Q2. Explain the concept of ϵ (epsilon)-insensitive loss in SVR.

.....

Q3. Differentiate between Linear Regression and Support Vector Regression.

.....

Q4. What are the advantages of Support Vector Regression?

.....

Q5. What are the limitations of Support Vector Regression?

.....

8.4 SUMMARY

Unit 8 on Regression Models provides a detailed understanding of one of the most important techniques in predictive analytics. It begins with the introduction to regression, which explains how relationships between variables are modeled to predict outcomes. Regression analysis focuses on identifying how a dependent variable changes with respect to one or more independent variables, making it a powerful tool for forecasting and decision-making in domains such as finance, healthcare, and marketing. The unit establishes the conceptual foundation by highlighting the importance of regression in understanding patterns and making data-driven predictions.

The unit then explores different types of regression models, starting with Linear Regression, which assumes a straight-line relationship between variables and is widely used due to its simplicity and interpretability. This is extended to Multiple Regression, where multiple independent variables are used to make predictions, reflecting real-world scenarios more accurately. When relationships are not linear, Polynomial Regression is introduced to model curved patterns, enabling better representation of complex data trends. These models collectively demonstrate how regression techniques can be adapted based on the nature of the dataset.

Further, the unit covers advanced techniques such as Ridge and Lasso Regression, which are used to address issues like overfitting and multicollinearity. These regularization methods improve model performance by controlling the magnitude of coefficients and, in the case of Lasso, even performing feature selection. Finally, the unit introduces Support Vector Regression (SVR), a powerful method capable of handling non-linear and complex relationships by fitting data within a specified margin of tolerance. Overall, the unit provides a comprehensive overview of regression models, equipping learners with both fundamental and advanced tools to build accurate and reliable predictive models.

8.5 ANSWERS

📌 Check Your Progress 1

Q1. What is Regression Analysis? Explain its purpose in predictive data analysis.

Solution: refer to Section 8.2

Q2. Explain the difference between dependent and independent variables with examples.

Solution: refer to Section 8.2

Q3. Write the general form of a regression equation and explain its components.

Solution: refer to Section 8.2

Q4. Discuss the challenges and limitations of regression analysis.

Solution: refer to Section 8.2

Q5. What is the role of regression in predictive analytics?

Solution: refer to Section 8.2

☛ Check Your Progress 2

Q1. What is a Regression Model? Explain its significance in predictive analytics.

Solution: refer to Section 8.3

Q2. Define Simple Linear Regression and write its mathematical equation.

Solution: refer to Subsection 8.3.1

Q3. Explain the Ordinary Least Squares (OLS) method used in linear regression.

Solution: refer to Subsection 8.3.1

Q4. Interpret the slope and intercept in a linear regression model..

Solution: refer to Subsection 8.3.1

Q5. List the assumptions and limitations of Simple Linear Regression

Solution: refer to Subsection 8.3.1

☛ Check Your Progress 3

Q1. What is Multiple Linear Regression? Explain its importance.

Solution: refer to Subsection 8.3.2

Q2. Write the general form of a Multiple Linear Regression model and explain its components.

Solution: refer to Subsection 8.3.2

Q3. Explain the interpretation of regression coefficients in Multiple Regression.

Solution: refer to Subsection 8.3.2

Q4. What is Multicollinearity? How is it detected?

Solution: refer to Subsection 8.3.2

Q5. Explain model selection techniques used in Multiple Regression.

Solution: refer to Subsection 8.3.2

☛ Check Your Progress 4

Q1. What is Polynomial Regression? Why is it used?.

Solution: refer to Subsection 8.3.3

Q2. Write the general form of a Polynomial Regression model.

Solution: refer to Subsection 8.3.3

Q3. How does Polynomial Regression differ from Linear Regression?

Solution: refer to Subsection 8.3.3

Q4. What are the advantages and limitations of Polynomial Regression?

Solution: refer to Subsection 8.3.3

Q5. When should Polynomial Regression be preferred over Linear Regression?

Solution: refer to Subsection 8.3.3

☛ Check Your Progress 5

Q1. What are Ridge and Lasso Regression? Why are they used?

Solution: refer to Subsection 8.3.4

Q2. Differentiate between Ridge Regression and Lasso Regression.

Solution: refer to Subsection 8.3.4

Q3. Explain the concept of regularization in regression.

Solution: refer to Subsection 8.3.4

Q4. How do Ridge and Lasso Regression help in handling multicollinearity?

Solution: refer to Subsection 8.3.4

Q5. What are the advantages and limitations of Ridge and Lasso Regression?

Solution: refer to Subsection 8.3.4

☛ Check Your Progress 6

Q1. What is Support Vector Regression (SVR)? Explain its objective.

Solution: refer to Subsection 8.3.5

Q2. Explain the concept of ϵ (epsilon)-insensitive loss in SVR.

Solution: refer to Subsection 8.3.5

Q3. Differentiate between Linear Regression and Support Vector Regression.

Solution: refer to Subsection 8.3.5

Q4. What are the advantages of Support Vector Regression?

Solution: refer to Subsection 8.3.5

Q5. What are the limitations of Support Vector Regression?

Solution: refer to Subsection 8.3.5

8.6 REFERENCES/FURTHER READINGS

1. *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, published by Springer (2nd Edition, 2021, ISBN: 978-1071614174).
2. *Introduction to Linear Regression Analysis* by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, published by Wiley (5th Edition, 2012, ISBN: 978-0470542811).
3. *Applied Regression Analysis and Generalized Linear Models* by John Fox, published by Sage Publications (3rd Edition, 2016, ISBN: 978-1452205663).
4. *Pattern Recognition and Machine Learning* by Christopher M. Bishop, published by Springer (1st Edition, 2006, ISBN: 978-0387310732).
5. *Data Mining: Concepts and Techniques* by Jiawei Han, Micheline Kamber, and Jian Pei, published by Morgan Kaufmann (3rd Edition, 2011, ISBN: 978-0123814791).
6. *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, published by Springer (2nd Edition, 2021, ISBN: 978-1071614174).
7. *Practical Statistics for Data Scientists* by Peter Bruce, Andrew Bruce, and Peter Gedeck, published by O'Reilly Media (2nd Edition, 2020, ISBN: 978-1492072942).
8. *Business Analytics: Data Analysis and Decision Making* by S. Christian Albright and Wayne L. Winston, published by Cengage Learning (7th Edition, 2020, ISBN: 978-0357131787).
9. *Introduction to Operations Research* by Frederick S. Hillier and Gerald J. Lieberman, published by McGraw-Hill Education (10th Edition, 2014, ISBN: 978-0073523453).
10. *Decision Analysis for Management Judgment* by Paul Goodwin and George Wright, published by Wiley (5th Edition, 2014, ISBN: 978-1119974017).
11. *Handbook of Statistical Analysis and Data Mining Applications* by Robert Nisbet, John Elder, and Gary Miner, published by Elsevier (2nd Edition, 2017, ISBN: 978-0124166455).
12. *Data Mining and Analysis: Fundamental Concepts and Algorithms* by Mohammed J. Zaki and Wagner Meira Jr., published by Cambridge University Press (1st Edition, 2014, ISBN: 978-0521766333).
13. *Core Concepts in Data Analysis: Summarization, Correlation and Visualization* by Boris Mirkin, published by Springer (1st Edition, 2011, ISBN: 978-0857292872).
14. *Applied Predictive Modeling*, Max Kuhn and Kjell Johnson, 1st Edition, Springer, 2013.
15. *An Introduction to Statistical Learning with Applications in R*, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 1st Edition, Springer, 2013.