PW SKILLS

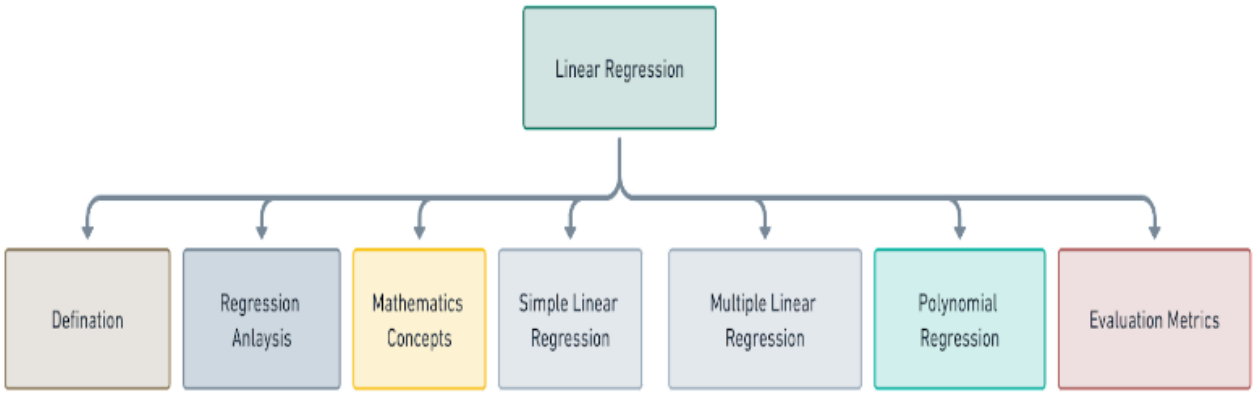
Lesson Plan

linear regression

Topics Covered

- Linear Regression
  - What is Regression analysis
  - The use of regression
  - Linear Regression(definition)
  - Assumptions of Linear Regression
  - Mathematics involved
  - Simple Linear Regression
  - Multiple Linear Regression
  - Polynomial Regression
  - Evaluation Metrics
  - Advantage and limitation of linear regression

Linear Regression :



What is Regression Analysis ?

Regression in statistics is the process of predicting a Label(or Dependent Variable) based on the features(Independent Variables) at hand. Regression is used for time series modelling and finding the causal effect relationship between the variables and forecasting. For example, the relationship between the stock prices of the company and various factors like customer reputation and company annual performance etc. can be studied using regression.

Regression analysis is an important tool for analysing and modelling data. Here, we fit a curve/line to the data points, in such a manner that the difference between the distance of the actual data points from the plotted curve/line is minimum.

The use of Regression :

Regression analyses the relationship between two or more features. Let's take an example:

Let's suppose we want to make an application which predicts the chances of admission to a foreign university. In that case, the

The benefits of using Regression analysis are as follows:

- It shows the significant relationships between the Label (dependent variable) and the features(independent variable).
- It shows the extent of the impact of multiple independent variables on the dependent variable.
- It can also measure these effects even if the variables are on a different scale.
- These features enable the data scientists to find the best set of independent variables for predictions.

Linear Regression :

Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with. The building blocks of a Linear Regression Model are:

- Discrete/continuous independent variables.
- A best-fit regression line.

Continuous dependent variable. i.e., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

$$Y=a+b*X + e$$

Where,

a is the intercept, b is the slope of the line, and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

Assumptions of Linear Regression : Linear regression relies on certain assumptions for accurate and reliable predictions:

Linearity: The relationship between the dependent variable and the independent variables is assumed to be linear. It means that the change in the dependent variable is directly proportional to the change in the independent variables.

Independence: The observations or data points are assumed to be independent of each other. In other words, the value of one observation does not depend on the value of another observation.

Homoscedasticity: The variance of the errors (residuals) is constant across all levels of the independent variables. Homoscedasticity indicates that the spread of the residuals is consistent across the range of predicted values.

Normality: The errors follow a normal distribution, meaning the residuals are normally distributed. Normality assumptions are essential for conducting hypothesis tests and constructing confidence intervals.

No perfect multicollinearity: There should be no perfect linear relationship between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated, making it challenging to distinguish their individual effects on the dependent variable.

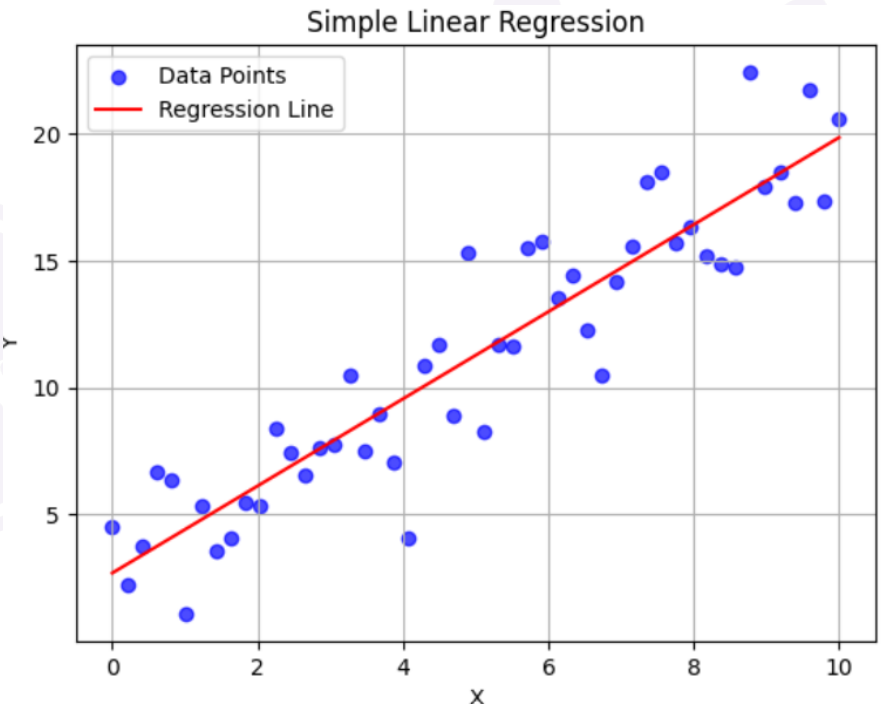
Simple Linear Regression : Simple Linear regression is a method for predicting a quantitative response using a single feature ("input variable"). The mathematical equation is:

y=mx+b

What do terms represent?

- is the response or the target variable
- is the independent variable
- m is the slope(coefficient) of line
- b is the intercept

m and c are the model coefficients. To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients.



Estimation of Coefficients in Simple Linear Regression :

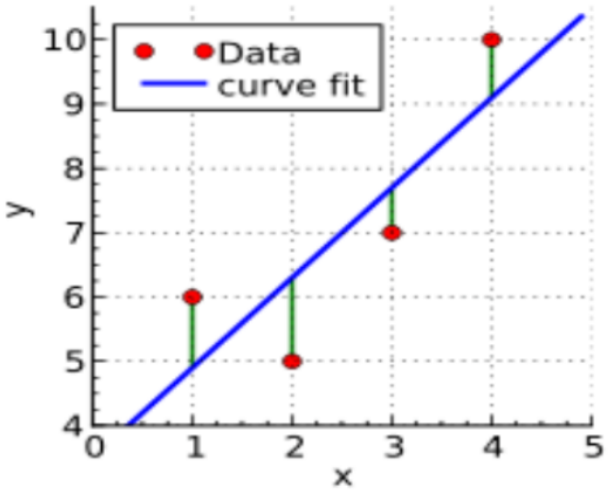
The coefficients (m and b) in the simple linear regression equation are estimated using the Ordinary Least Squares (OLS) method. OLS finds the coefficients that minimise the sum of squared differences between the observed and predicted values.

The steps involved in estimating the coefficients are:

- Calculate the mean of the independent variable (x) and the mean of the dependent variable (y).
- Calculate the covariance between x and y.
- Calculate the variance of x.
- Calculate the slope m as the ratio of the covariance to the variance of x.
- Calculate the intercept b using the slope m and the means of x and y.

The mathematics involved :

Take a quick look at the plot created. Now consider each point, and know that each of them has a coordinate in the form (X, Y). Now draw an imaginary line between each point and the current "best-fit" line. We'll call the distance between each point and the current best-fit line as D. To get a quick image of what we're trying to visualise, take a look at the picture below.



What elements are present in the diagram ?

- The red points are the observed values of x and y.
- The blue line is the least squares line.
- The green lines are the residuals, which is the distance between the observed values and the least squares line.

Before, we're labelling each green line as having a distance D, and each red point as having a coordinate of (X, Y). Then we can define our best fit line as the line having the property were:

D12+D22+D32+.....+Dn2

So how do we find this line? The least-square line approximating the set of points:

$$(X, Y)_1, (X, Y)_2, (X, Y)_3, (X, Y)_4, (X, Y)_5,$$

Has the equation :

Y= a0+a1*X

This is basically just a rewritten form of the standard equation for a line:

Y = mx+b

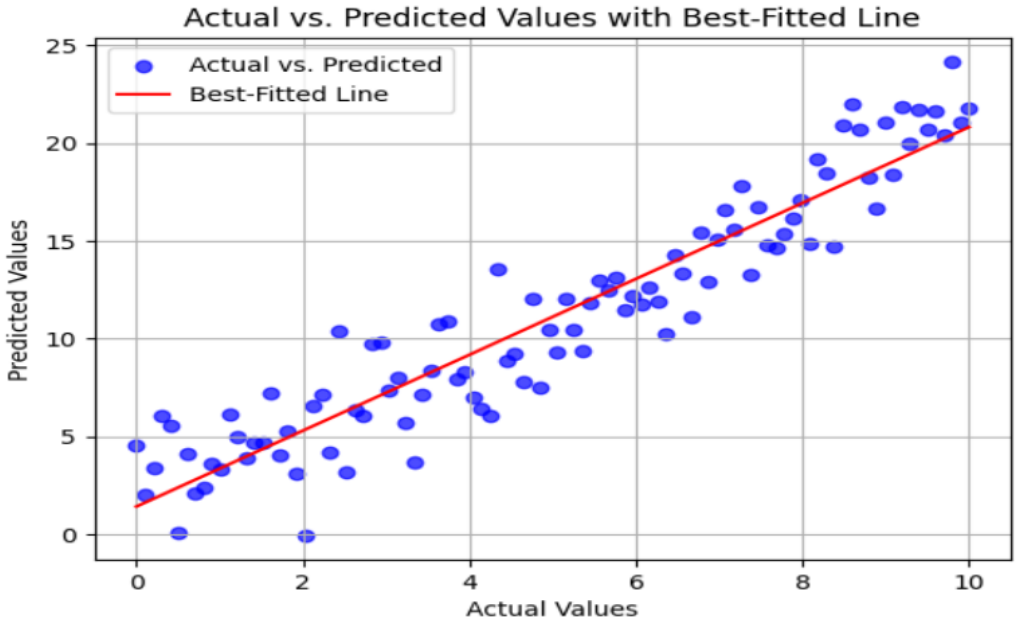
We can solve for these constants a0 and a1 by simultaneously solving these equations:

$$\Sigma Y = a_0 N + a_1 \Sigma X$$
$$\Sigma XY = a_0 \Sigma X + a_1 \Sigma X^2$$

These are called the normal equations for the least-squares line. There are further steps that can be taken in rearranging these equations to solve for y, but we'll let scikit-learn do the rest of the heavy lifting here.

How do you know this is the best fit line ?

The best fit line is obtained by minimising the residual. Residual is the distance between the actual Y and the predicted Y:



Mathematically, Residual is r = y-(mx+b)

Hence the sum of square of residual is :

$$r_i = y_i - (mx_i + b) \quad \text{(Residual for one point)}$$

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} (y_i - (mx_i + b)) \quad \text{(Sum of residuals)}$$

$$R(x) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (mx_i + b))^2 \quad \text{(Sum of squares of residuals)}$$

As we can see that the residual is both a function of m and b, so differentiating partially with respect to m and b will give us.

$$\frac{\partial R}{\partial m} = \sum_{i=0}^{n} 2x_i (b + mx_i - y_i)$$

$$\frac{\partial R}{\partial b} = \sum_{i=0}^{n} 2 (b + mx_i - y_i)$$

For getting the best fit line, residual should be minimum. The minima of a function occurs where the derivative=0. So, equating our corresponding derivatives to 0, we get.

$$\sum_{i=0}^{n} 2x_i (b + mx_i - y_i) = 0$$

$$\sum_{i=0}^{n} 2 (b + mx_i - y_i) = 0$$

$$-$$

$$\sum_{i=0}^{n} 2x_i b + 2mx_i^2 - 2y_i x_i = 0$$

$$\sum_{i=0}^{n} 2b + 2mx_i - 2y_i = 0$$

$$-$$

$$\sum_{i=0}^{n} 2x_i b + \sum_{i=0}^{n} 2mx_i^2 - \sum_{i=0}^{n} 2y_i x_i = 0$$

$$\sum_{i=0}^{n} 2b + \sum_{i=0}^{n} 2mx_i - \sum_{i=0}^{n} 2y_i = 0 \quad \text{(Break up the summations)}$$

$$-$$

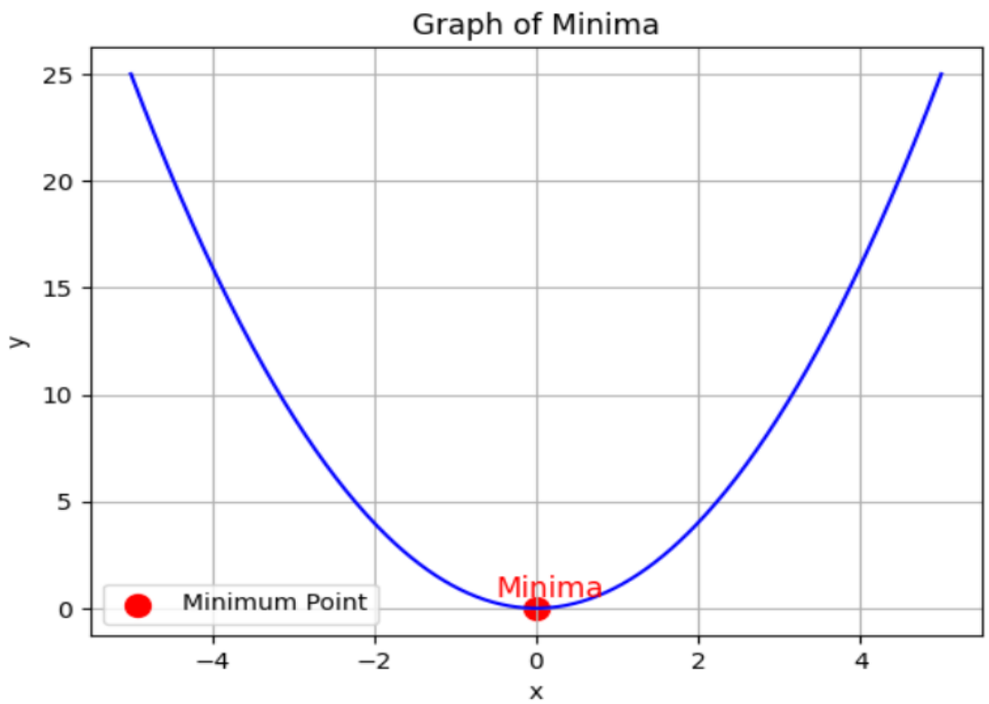
$$\sum_{i=0}^{n} x_i b + \sum_{i=0}^{n} mx_i^2 - \sum_{i=0}^{n} y_i x_i = 0$$

$$\sum_{i=0}^{n} b + \sum_{i=0}^{n} mx_i - \sum_{i=0}^{n} y_i = 0 \quad \text{(diving both sides by 2)}$$

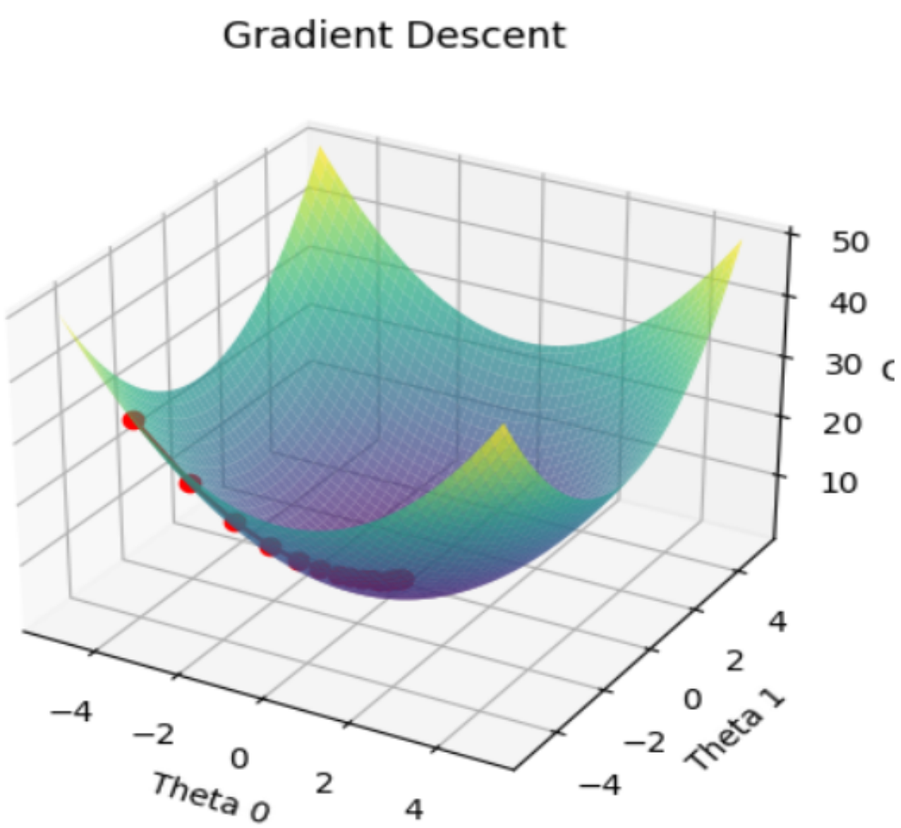
This same equation can be written in matrix form as

$$\begin{bmatrix} \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \\ n & \sum_{i=0}^{n} x_i \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} y_i x_i \\ \sum_{i=0}^{n} y_i \end{bmatrix}$$

Ideally, if we'd have an equation of one dependent and one independent variable the minima will look as follows:



But as the residual minima is dependent on two variables m and b, it becomes a Paraboloid and the appropriate m and b are calculated using Gradient Descent as shown below:



Now, let's understand how to check how well the model fits our data.

The new values for 'slope' and 'intercept' are calculated as follows:

$$\begin{aligned}&\text{repeat until convergence } \{\\&\quad \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)\\&\quad \theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) \cdot x^{(i)}\\&\}\end{aligned}$$

where, 0 is 'intercept' , 1is the slope, is the learning rate, m is the total number of observations and the term after the Σsign is the loss. Google Tensor board recommends a Learning rate between 0.00001 and 10. Generally a smaller learning rate is recommended to avoid overshooting while creating a model.

R Square Statistics :

The R-squared statistic provides a measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1. In simple words, it represents how much of our data is being explained by our model. For example, 2 statistic = 0.75, it says that our model fits 75 % of the total data set. Similarly, if it is 0, it means none of the data points is being explained and a value of 1 represents 100% data explanation. Mathematically 2 statistic is calculated as :

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

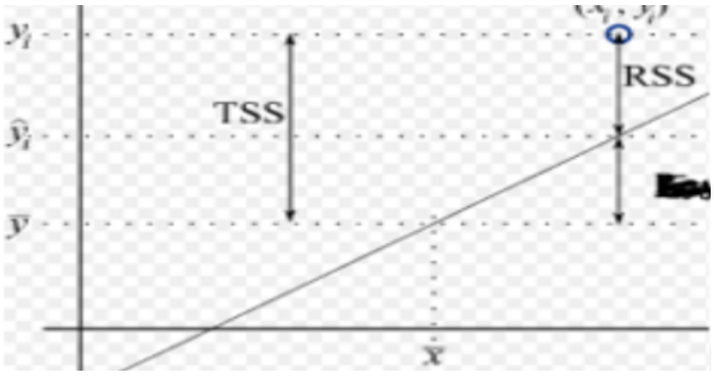
Where RSS: is the Residual Sum of squares and is given as :

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

RSS is the residual(error) term we have been talking about so far. And, TSS: is the Total sum of squares and given as :

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

TSS is calculated when we consider the line passing through the mean value of y, to be the best fit line. Just like RSS, we calculate the error term when the best fit line is the line passing through the mean value of y and we get the value of TSS.



The closer the value of R2 is to 1 the better the model fits our data. If R2 comes below 0(which is a possibility) that means the model is so bad that it is performing even worse than the average best fit line.

Adjusted R squared : As we increase the number of independent variables in our equation, the R2 increases as well. But that doesn't mean that the new independent variables have any correlation with the output variable. In other words, even with the addition of new features in our model, it is not necessary that our model will yield better results but R2 value will increase. To rectify this problem, we use Adjusted R2 value which penalises excessive use of such features which do not correlate with the output data.

$$R^2 \text{ adjusted} = 1 - \frac{(1-R2)(N-1)\,(N-1)}{N-p-1}$$

Where,

R2 = Sample R-Square
p = Number for predictors
N = Total Sample Size

In the equation above, when p = 0, we can see that adjusted R2 becomes equal to R2. Thus, adjusted R2 will always be less than or equal to R2, and it penalises the excess of independent variables which do not affect the dependent variable.

Multiple Linear Regression :

Multiple linear regression is an extension of simple linear regression that involves two or more independent variables to predict a single dependent variable. It allows us to model the relationship between the dependent variable and multiple predictors simultaneously. Multiple linear regression is a widely used statistical method in various fields, including economics, social sciences, and data science.

Multiple Linear Regression Equation :

In multiple linear regression, the relationship between the dependent variable y and n independent variables (x1, x2, ..., xn) can be represented by the following equation:

y=b0+b1*x1+b2*x2+b3*x3+.....+bn*xn

Where :

- y is the dependent variable,
- x1, x2, ..., xn are the independent variables,
- b0 is the intercept, and
- b1, b2, ..., bn are the coefficients for the respective independent variables.

Each coefficient bi represents the change in the dependent variable y for a one-unit change in the corresponding independent variable xi, while keeping all other independent variables constant. The intercept b0 represents the value of the dependent variable y when all the independent variables are zero

Estimation of Coefficients in Multiple Linear Regression :

The coefficients (b0, b1, b2, ..., bn) in multiple linear regression are estimated using the Ordinary Least Squares (OLS) method. OLS finds the coefficients that minimise the sum of squared differences between the observed and predicted values.

The steps involved in estimating the coefficients are similar to simple linear regression:

- Calculate the mean of each independent variable and the mean of the dependent variable.
- Calculate the covariance matrix between the dependent variable and each independent variable.
- Calculate the variance-covariance matrix of the independent variables.

Calculate the coefficient vector b using the formula b = (XT * X)-1 * XT * y, where X is the matrix of independent variables and y is the vector of the dependent variable.

Advantages of Multiple Linear Regression :

Modelling Complex Relationships: Multiple linear regression can capture more complex relationships between the dependent variable and multiple predictors. It allows for the simultaneous examination of the effects of multiple variables on the outcome.

Interpretability: While the interpretation becomes more complex with multiple predictors, the coefficients can still provide valuable insights into the relationships between the variables.

Feature Importance: The coefficients can help determine the relative importance of each predictor in predicting the dependent variable.

Limitations of Multiple Linear Regression :

Assumptions: Multiple linear regression assumes linearity, independence, homoscedasticity, normality, and no perfect multicollinearity. Violations of these assumptions can lead to biassed or unreliable results.

Overfitting: As with any regression method, including irrelevant or highly correlated predictors can lead to overfitting, where the model performs well on the training data but poorly on unseen data.
Multicollinearity: When independent variables are highly correlated, it can be challenging to determine their individual effects on the dependent variable.

Polynomial Regression :

Polynomial regression is an extension of linear regression that allows us to capture nonlinear relationships between the dependent variable and the independent variables. Instead of fitting a straight line, polynomial regression fits a polynomial curve to the data. It is particularly useful when the linear relationship does not accurately represent the underlying data.

Polynomial Regression Equation : In polynomial regression, the relationship between the dependent variable y and the independent variable x is modelled using a polynomial equation of the form.

y=b0+b1*x+b2*x2+.....+bn*xn

Where,

- y is the dependent variable.
- x is the independent variable.
- b0, b1, b2, ..., bn are the coefficients, and
- n is the degree of the polynomial.

The degree n determines the flexibility of the curve. A linear regression corresponds to a polynomial regression with n = 1, while a quadratic regression has n = 2, and so on. The higher the degree, the more complex the curve, and the better the model can capture nonlinear patterns.

Advantages of Polynomial Regression :

Flexibility: Polynomial regression can capture more complex relationships between variables than linear regression. It allows for a more flexible model that can fit curved data patterns.

Better Fit: In situations where the underlying relationship is nonlinear, polynomial regression can provide a better fit to the data compared to linear regression.

Easy to Implement: Polynomial regression is an extension of linear regression, so the same principles and techniques can be used for model evaluation and interpretation.

Limitations of Polynomial Regression :

Overfitting: As the degree of the polynomial increases, the model becomes more flexible and can closely fit the training data. However, a very high degree can lead to overfitting, where the model captures noise and random fluctuations in the data, leading to poor generalisation to unseen data.

Interpretability: As the complexity of the polynomial increases, the interpretability of the model decreases. Higher-degree polynomials can be challenging to interpret and visualise.

Extrapolation: Polynomial regression may not provide accurate predictions outside the range of the training data, especially when using high-degree polynomials. Extrapolation can lead to unreliable predictions.

Polynomial Regression vs. Linear Regression:

The choice between polynomial regression and linear regression depends on the underlying relationship between the variables and the nature of the data. Linear regression is suitable when there is a linear relationship between the variables, whereas polynomial regression is more appropriate when the relationship is nonlinear and cannot be adequately represented by a straight line.

Model Evaluation :

After fitting the linear regression model, it's crucial to evaluate its performance. Common metrics for evaluating linear regression models include:

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values. It quantifies the average magnitude of errors, giving more weight to large errors.

Root Mean Squared Error (RMSE): The square root of MSE, providing a more interpretable error metric. RMSE is in the same unit as the dependent variable, making it easier to understand the model's prediction errors.

R-squared (R2): Represents the proportion of variance in the dependent variable that is predictable from the independent variables. R2 value ranges from 0 to 1, with 1 indicating a perfect fit. It is a measure of how well the independent variables explain the variability in the dependent variable. Higher R2 values indicate a better fit, but it's essential to interpret R2 in the context of the problem domain. It is possible to achieve a high R2 value even when the model is not practically useful or when there is overfitting.

Adjusted R2 : Regression analysis uses adjusted R-squared, a statistical metric, to evaluate the goodness of fit of a regression model. It is a development of the standard (unadjusted) R-squared, which measures the percentage of the dependent variable's variation that can be accounted for by the model's independent variables.

Advantages and Limitations of Linear Regression :

The advantages of linear regression include its simplicity, interpretability, and ease of implementation. It is a powerful tool for modelling relationships between variables in a straightforward manner. Additionally, linear regression can be useful for generating insights into the direction and strength of relationships between variables.However, linear regression has limitations:

Sensitivity to Outliers: Linear regression can be sensitive to outliers, which are data points that significantly deviate from the overall pattern. Outliers can have a substantial impact on the slope and intercept of the regression line.

Assumption of Linearity: Linear regression assumes a linear relationship between the dependent and independent variables. In reality, many relationships may not be strictly linear, which can lead to inaccurate predictions.

Limited to Linear Relationships: As the name suggests, linear regression is limited to modelling linear relationships. It may not capture the complex nonlinear relationships present in some datasets.

Overfitting: If the model becomes too complex (e.g., by including too many irrelevant or highly correlated predictors), it can lead to overfitting, where the model performs well on the training data but poorly on unseen data.