Data Science Learning Accelerator

Supervised Machine Learning Algorithms

Understanding Linear Regression

Fall 2023

Sravya Vujjini

Linear Regression

What is Linear Regression?

Linear Regression is a fundamental statistical and supervised machine learning technique used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors or features) by fitting a linear equation. It is one of the simplest and most widely used regression techniques, and its primary goal is to find the best-fitting line, known as the regression line, that represents the relationship between the variables.

The two main types of Linear Regression are:

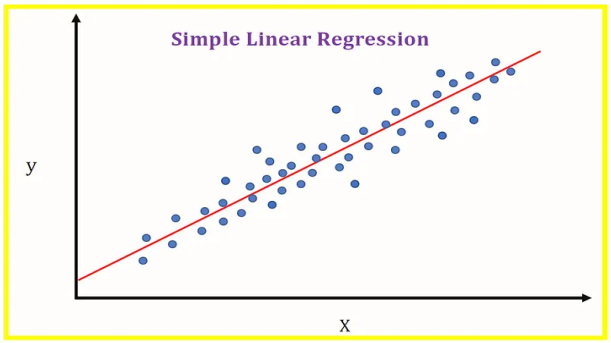
1. Simple Linear Regression 2. Multiple Linear Regression

Simple Linear Regression

In the Simple Linear Regression, there is only one independent variable (X) and one dependent variable (Y), and the relationship is expressed as:

$$Y = a + bX$$

- Y represents the dependent variable (target), which you want to predict or understand.
- X represents the independent variable (predictor), which influences Y.
- "a" is the intercept, indicating the predicted value of Y when X is 0.
- "b" is the slope, representing the change in Y for a one-unit change in X.



In the above figure, you can observe the best possible line to predict the value of Y based on X.

How does linear regression determine the optimal fit line?

In linear regression, the objective is to determine the most suitable values for a (intercept) and b (slope) in order to identify the ideal fit line. This ideal line minimizes the error, ensuring that the error between predicted and actual values is as low as possible.

Step-by-step explanation of how the best-fit line is determined in simple linear regression

1.Scatter Plot: Begin by creating a scatter plot of your data, where the x-axis represents the independent variable (X), and the y-axis represents the dependent variable (Y). This plot helps visualize the relationship between the variables.

2.Visualize the Line: Initially, you may draw a straight line on the scatter plot that you believe represents the relationship between X and Y. This line is just an initial estimate.

3.Calculate Residuals: For each data point, calculate the vertical distance (residual) between the actual Y value and the value predicted by your initial line. These residuals represent the errors in your initial estimate.

4.Mean Squared Error (MSE): To evaluate the quality of your initial line, compute the Mean Squared Error (MSE) by squaring each residual, summing up these squared errors, and then dividing by the number of data points.

5.Minimization: The objective in linear regression is to minimize the MSE. This is done by finding the values of 'a' and 'b' that lead to the smallest MSE. These values represent the intercept and the slope of the best-fit line.

6.Plot the Best-Fit Line: The values of 'a' and 'b' that you've calculated represent the intercept and the slope of the best-fit line. Draw the best-fit line minimizes the Mean Squared Error (MSE) and provides the most accurate representation of the relationship between X and Y in terms of a linear equation on the scatter plot.

Multiple Linear Regression

Multiple linear regression extends the concept of simple linear regression by incorporating multiple independent variables (predictors) to predict a dependent variable (target).The relationship in multiple linear regression is expressed as:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

- Y represents the dependent variable you aim to predict or understand.
- $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ are the independent variables that influence Y.
- a is the intercept, indicating the predicted value of Y when all independent variables $(X_1, X_2, \ldots, X_n)$ are zero.
- $\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_n}$ are the slopes, representing the change in Y for a one-unit change in each of the respective independent variables $(X_1, X_2, \ldots, X_n)$.

The goal in multiple linear regression is to determine the best-fitting line in a multidimensional space, where each independent variable contributes to predicting the dependent variable. The principles of minimizing the error, using methods like Mean Squared Error (MSE) remain the same, but the calculations become more complex with multiple predictors.

Advantages and Disadvantages of Linear Regression

Advantages :

- **Simplicity:** Linear regression is straightforward and easy to understand. It provides a simple and interpretable model.
- **Interpretability:** The coefficients (slopes and intercept) of the linear model have clear and intuitive interpretations. This makes it easy to explain the relationships between variables to non-technical audiences.

- **Speed:** Linear regression models are computationally efficient and can handle large datasets relatively quickly.
- **Useful for Initial Analysis:** Linear regression can serve as a baseline model for many problems. It's a useful starting point for understanding the data and relationships between variables.

Disadvantages :

- **Linearity Assumption:** Linear regression assumes a linear relationship between the dependent and independent variables. This assumption may not hold for many real-world problems, leading to model inadequacy.
- **Limited Expressiveness:** Linear regression is limited in its ability to capture complex relationships between variables. It may not perform well for non-linear data.
- **Sensitive to Outliers:** Linear regression can be sensitive to outliers, leading to biased parameter estimates and model performance.
- **Independence Assumption:** Linear regression assumes that the errors (residuals) are independent. Violation of this assumption can result in inaccurate parameter estimates.
- **Multicollinearity:** When independent variables are highly correlated (multicollinearity), it can be challenging to separate their individual effects on the dependent variable.
- **Overfitting and Underfitting:** Linear regression may suffer from overfitting or underfitting, depending on the complexity of the model and the quality of the data.

- **Data Transformation:** For non-linear relationships, you may need to transform the data or use polynomial regression, which can be more complex.

References :

- OpenAI. (2023). ChatGPT (Sep 25 version) [Large language model]. https://chat.openai.com/chat.
- https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/h-what-is-linear-regression