Logistic regression is a supervised learning algorithm used for classification tasks, where the goal is to predict the probability of an outcome belonging to one of two classes (binary classification). Unlike linear regression, which predicts continuous values, logistic regression outputs probabilities, which can be mapped to class labels.

Core Concepts of Logistic Regression

1. **Logistic (Sigmoid) Function:**
   - The core of logistic regression is the sigmoid function, which maps any real-valued number to a value between 0 and 1.
   - For a given input x, the logistic function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

   - This function helps to model probabilities that our output belongs to one of two classes.
2. **Probability and Thresholding:**
   - Logistic regression calculates the probability of a sample belonging to a certain class (e.g., $P(Y = 1|X)$).
   - Typically, we set a threshold (often 0.5) to classify observations: if $P(Y = 1|X) > 0.5$, classify as 1; otherwise, classify as 0.
3. **Log-Odds and Linear Relationship:**
   - Logistic regression models the log-odds (logarithm of the odds ratio) of the probability of the target class as a linear combination of the input features.
   - The equation is:

$$\ln\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

   - This log-odds formulation maintains a linear relationship between the features and the transformed probability, allowing the model to learn from a linear structure.

Assumptions of Logistic Regression

1. **Linearity of Independent Variables and Log-Odds:**
   - Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. This doesn't mean the data itself has to be linear but rather that the effect of each predictor on the log-odds is linear.
   - Polynomial or interaction terms can be introduced if the relationship is not naturally linear.
2. **Independent Observations:**
   - Observations should be independent of each other. In other words, the data should not contain duplicated or strongly correlated entries.

- Violation of this assumption, such as in time-series data with autocorrelation, can lead to biased estimates.

3. **No Multicollinearity Among Predictors:**
   - High correlation between predictor variables (multicollinearity) can distort the relationship between predictors and the outcome, making it hard to determine the independent effect of each predictor.
   - Checking for multicollinearity (using variance inflation factors, for example) is a good practice.
4. **Binary Outcome Variable:**
   - Logistic regression is inherently designed for binary classification tasks. However, it can be extended to multinomial and ordinal logistic regression for multiple classes.
5. **Large Sample Size:**
   - Logistic regression benefits from a large sample size to produce stable and reliable estimates. Small sample sizes can lead to overfitting or underfitting, especially with many predictors.
6. **Absence of Strong Outliers:**
   - Outliers in predictor variables can have a disproportionately large influence on the model and may skew results.
   - Outliers should ideally be removed or adjusted, or a more robust method should be considered if outliers are expected to impact the results.
7. **Appropriate Data Scaling:**
   - Although logistic regression does not require strict normalization, data scaling can improve performance and training stability, especially when regularization is applied.

Use Cases of Logistic Regression

- **Binary Classification Problems:** Predicting outcomes that have two categories, such as spam/not spam, default/no default, and disease/no disease.
- **Probability Estimation:** Besides classification, logistic regression gives the probability of each class, which can inform decision-making in scenarios like risk analysis.