# Logistic Regression

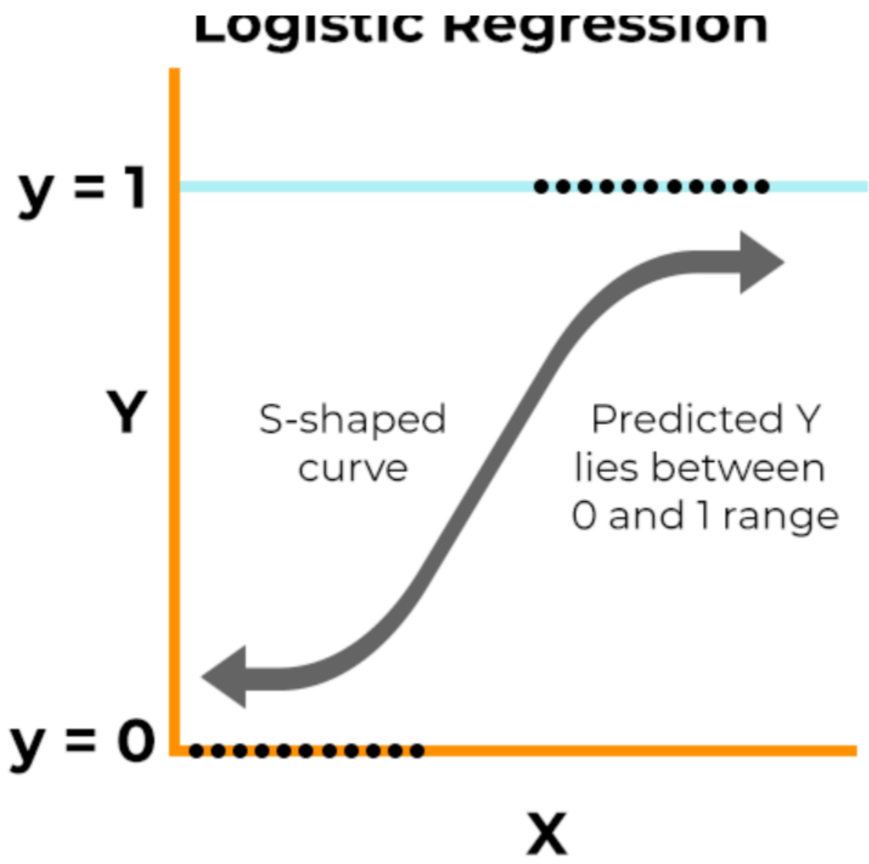
## 1. Introduction

Logistic regression is a statistical technique used for **classification** problems where the dependent variable (response) is **categorical** (e.g., yes/no, 0/1).
Instead of fitting a straight line, it models the **probability** that a given observation belongs to a particular class.

- Dependent variable (Y): categorical outcome (binary or multinomial).
- Independent variable (X): predictor(s), explanatory variables.

Logistic regression is used for:

- Classification
- Probability estimation
- Hypothesis testing
- Modeling binary outcomes (e.g., disease vs no disease, success vs failure).




---

## 2. The Logistic Regression Model

For binary outcomes ($Y \in \{0,1\}$):

$P(Y=1 \mid X) = 1 / (1 + e^{-(\beta_0 + \beta_1 X)})$

- $P(Y=1 \mid X)$: probability that Y=1 given predictor X.
- $\beta_0$: intercept.
- $\beta_1$: slope, effect of X on the log-odds of Y.

**Logit Transformation (Linear form):**

$\log\left( P(Y=1 \mid X) / (1 - P(Y=1 \mid X)) \right) = \beta_0 + \beta_1 X$

This makes the relationship between predictors and the log-odds linear.

---

## 3. Example

Suppose the fitted logistic regression equation is:

log( p / (1 - p) ) = -3 + 0.2X

- Interpretation of slope ($\beta 1 = 0.2$): For every unit increase in X, the log-odds of Y=1 increase by 0.2.
- Convert to probability:
  If X = 10, then

  p = 1 / (1 + e^-(-3 + 0.2*10))
  = 1 / (1 + e^1)
  ≈ 0.27

So the probability of Y=1 is about 27%.

---

## 4. Estimating Parameters (Maximum Likelihood)

Unlike linear regression, logistic regression uses **Maximum Likelihood Estimation (MLE)** to estimate coefficients.

- Define likelihood of observing given data under parameters β.
- Choose parameters that maximize this likelihood.
- Solved using iterative methods (e.g., Newton-Raphson, gradient descent).

---

## 5. Predicted and Residual Values

- Predicted value (p-hat): estimated probability of Y=1 for observation i.
- Residuals: difference between actual outcome and predicted probability.
    - Deviance residuals are often used.

---

## 6. Model Evaluation

Because logistic regression predicts probabilities, evaluation metrics differ from linear regression.

- Accuracy
- Confusion Matrix (TP, TN, FP, FN)
- Precision, Recall, F1-score
- ROC Curve and AUC
- Log-Loss (Cross-Entropy Loss)
- Pseudo $R^2$ (e.g., McFadden's $R^2$)

---

## 7. Multiple Logistic Regression

Extension of simple logistic regression to multiple predictors:

$\log( p / (1 - p) ) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

- Each $\beta_i$: effect of predictor $x_i$ on log-odds of Y=1, holding other predictors constant.

---

## 8. Assumptions of Logistic Regression

1. The outcome is binary (or categorical for multinomial).
2. Observations are independent.
3. No multicollinearity among predictors.
4. Large sample size preferred for stable estimates.
5. Relationship between predictors and log-odds is linear.

---

## 9. Applications

- Medical diagnosis (disease vs no disease).
- Credit scoring (default vs non-default).
- Marketing (buy vs not buy).
- Natural language processing (spam vs not spam).