Logistic Regression for Binary Classification

Rahul Shandilya

Introduction

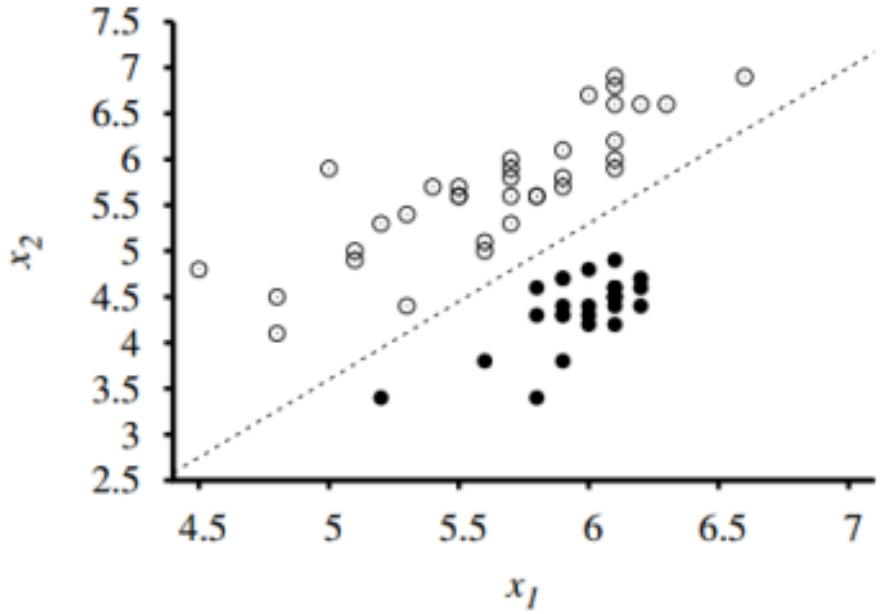
Definition

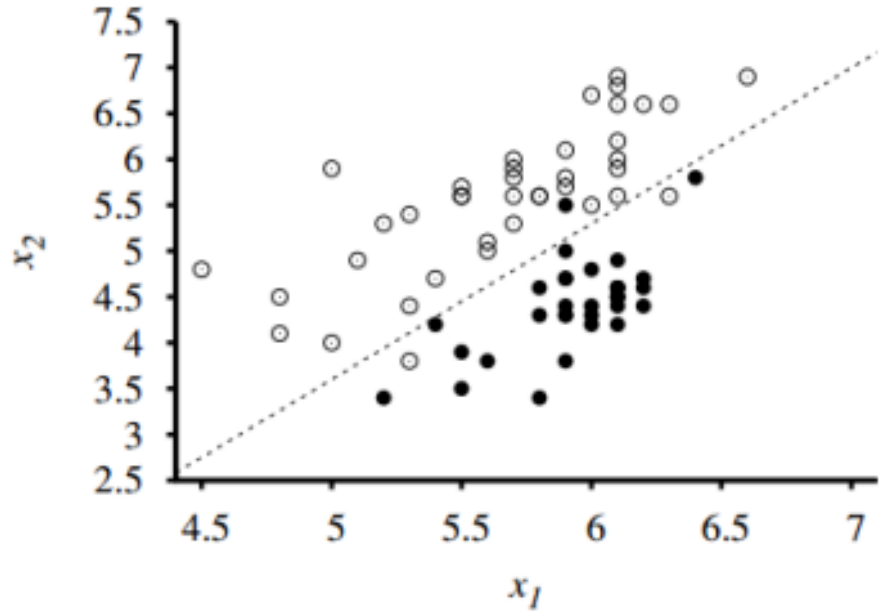
Logistic regression is a specialized form of regression that is formulated to predict and explain a binary (two-group) categorical variable rather than a metric dependent measure.

- It is the appropriate statistical technique when the dependent variable is a categorical (nominal or nonmetric) variable and the independent variables are metric or nonmetric variables.
- Logistic regression has widespread application in situations in which the primary objective is to identify the group to which an object (e.g., person, firm, or product) belongs.
- Potential applications include predicting anything where the outcome is binary (e.g., Yes/No). Such situations include the success or failure of a new product, deciding whether a person should be granted credit, or predicting whether a firm will be successful.
- In each instance, the objects fall into one of two groups, and the objective is to predict and explain the bases for each object's group membership through a set of independent

Example

Figure a shows data points of two classes: earthquakes (which are of interest to seismologists) and underground explosions (which are of interest to arms control experts). Each point is defined by two input values, x_1 and x_2, that refer to body and surface wave magnitudes computed from the seismic signal. Given these training data, the task of classification is to learn a hypothesis h that will take new $(x_1; x_2)$ points and return either 0 for earthquakes or 1 for explosions.





Logistic Regression Model

Even though the response may be a two outcome qualitative variable, we can always code the two cases as 0 and 1. For instance, we can take male $= 0$ and female $= 1$. Then the probability p of 1 is a parameter of interest. It represents the proportion in the population who are coded 1.

Linear Model

Let the response Y be either 0 or 1. If we were to model the probability of 1 with a single predictor linear model, we would write

$$p = E(Y|z) = \beta_0 + \beta_1 z$$

and then add an error term ϵ. But there are serious drawbacks to this model.

- The predicted values of the response Y could become greater than 1 or less than 0 because the linear expression for its expected value is unbounded.
- One of the assumptions of a regression analysis is that the variance of Y is constant across all values of the predictor

Logit Model

- Let us consider the *odds ratio*

$$odds = \frac{p}{1-p}$$

  which is the ratio of the probability of 1 to the probability of 0. Note, unlike probability, the odds ratio can be greater than 1.
- In logistic regression for a binary variable, we model the natural log of the odds ratio, which is called $logit(p)$.

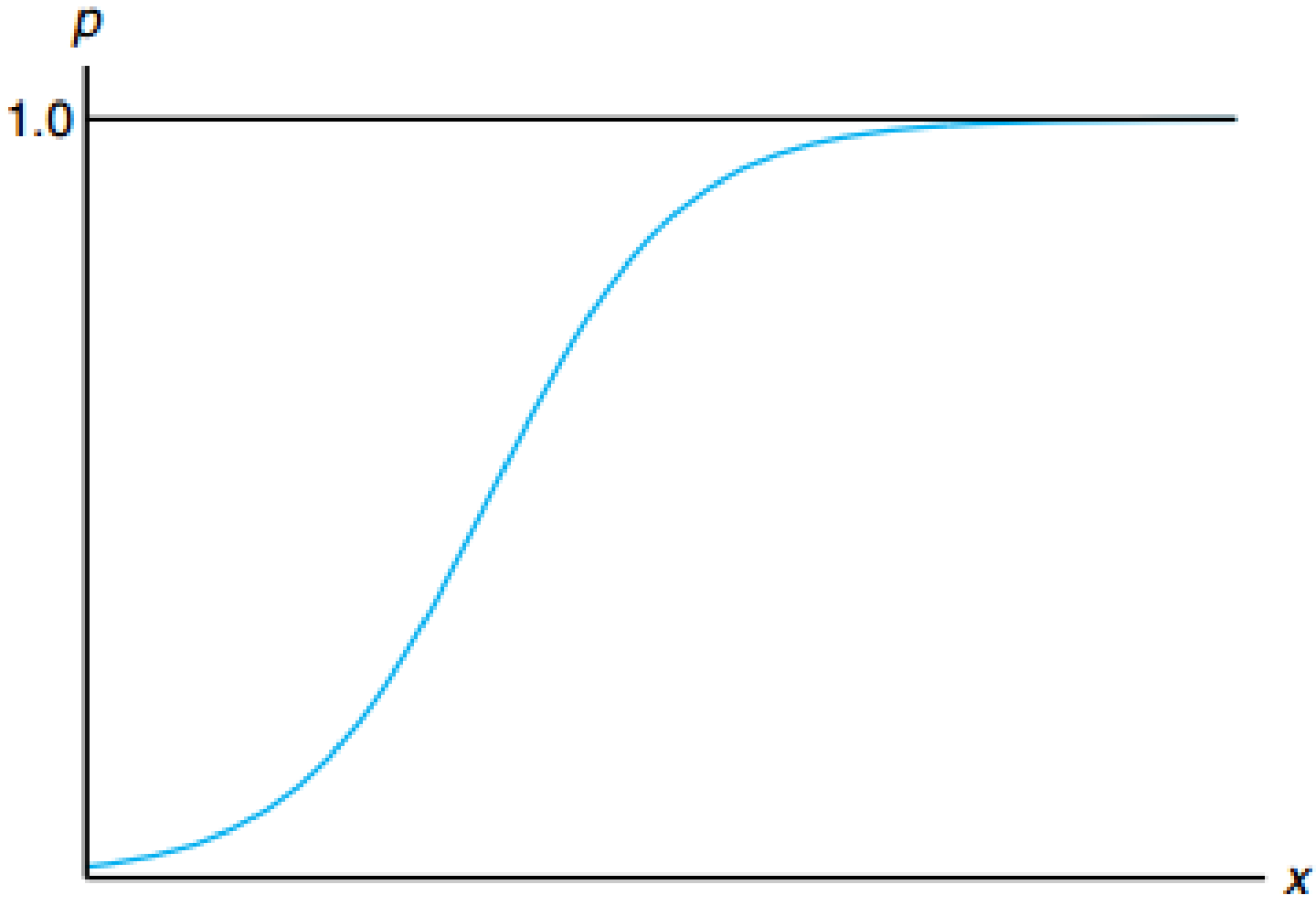
$$logit(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 z$$

- Solving for p, the logistic curve can be written as

$$p(z) = \frac{1}{1 + exp(-\beta_0 - \beta_1 z)}$$

- The relation between p and the predictor z is not linear but has an S-shaped graph. The parameter in the logistic curve determines how quickly p changes with z but its interpretation is not as simple as in ordinary linear regression because the relation is not linear, either in z or β_1.

Logistic Curve



Multivariate Logistic regression

- Let $(z_{j1}, z_{j2}, \ldots, z_{jr})$ be the values of the r predictors for the j-th observation. Setting first entry equal to 1 like regression analysis we have $\mathbf{z}_j = [1, z_{j1}, z_{j2}, \ldots, z_{jr}]'$
- we assume that the observation Y_j is Bernoulli with success probability $p(\mathbf{z_j})$depending on the values of the covariates. Then

$$\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) = \beta_0 + \beta_1 z_1 + \ldots + \beta_r z_r = \beta' \mathbf{z_j}$$

  where $\beta = [\beta_0 + \beta_1 + \ldots + \beta_r]'$
- Estimates of the β can be obtained by the method of maximum likelihood. The likelihood L is given by the joint probability distribution evaluated at the observed counts y_j.