Logistic Regression

Epoch IIT Hyderabad

Himani Agrawal
MA22BTECH11008

Contents

Introduction

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. The central mathematical concept that underlies logistic regression is the logit—the natural logarithm of an odds ratio.

Formulas

Simple logistic regression model has a form

$$\text{logit}(Y)=\ln(p/(1-p))=a+bx$$

Where, p=Probability(Y = outcome of interest | X = x, a specific value of X)

$$p=e^{a+bx}/(1+e^{a+bx})$$

Extending simple logistic regression to multiple we would have,

$$\text{logit}(Y)=\ln(p/(1-p))=a+b_1x_1+b_2x_2$$

p = Probability (Y = outcome of interest | $X_1 = x_1$, $X_2 = x_2$)

$$p=e^{a+b1x1+b2x2}/(1+e^{a+b1x1+b2x2})$$

Assumptions

1. Logistic regression requires the dependent variable to be discrete mostly dichotomous.
2. Since logistic regression estimates the probability of the event occurring (P(Y=1)), it is necessary to code the dependent variable accordingly. That is the desired out-come should be coded to be 1.
3. The model should be fitted correctly. It should not be over fitted with the meaningless variables included. Also it should not be under fitted with meaningful variable not included.

4. Logistic regression requires each observation to be independent. Also the model should have little or no multicollinearity. That is, independent variables are not linear functions of each other.
5. Whilst logistic regression does not require a linear relationship between the dependent and independent variables, it requires that the independent variables are linearly related to the log odds of an event.

Model Fitting

For the model fit the coefficients could be estimated using maximum likelihood estimation,which is achieved by maximizing likelihood function,binary logistic regressions could be calculated by iteratively reweighted least squares(IRLS).other methods include using bayesian statistics and rule of ten.