Logistic Regression

It is used to predict the result of a **categorical** dependent variable based on one or more continuous or categorical independent variables. In other words, it is multiple regression analysis but with a dependent variable is categorical.

Examples

1. An employee may get promoted or not based on age, years of experience, last performance rating etc. We are trying to calculate the factors that affects promotion. In this case, two possible categories in dependent variable : "**Promoted**" and "**Not Promoted**".

Algorithm

Logistic regression is based on Maximum Likelihood (ML) Estimation which says coefficients should be chosen in such a way that it maximizes the Probability of Y given X (likelihood). With ML, the computer uses different "iterations" in which it tries different solutions until it gets the maximum likelihood estimates. Fisher Scoring is the most popular iterative method of estimating the regression parameters.

logit(p) = b0 + b1X1 + b2X2 + ------ + bk Xk

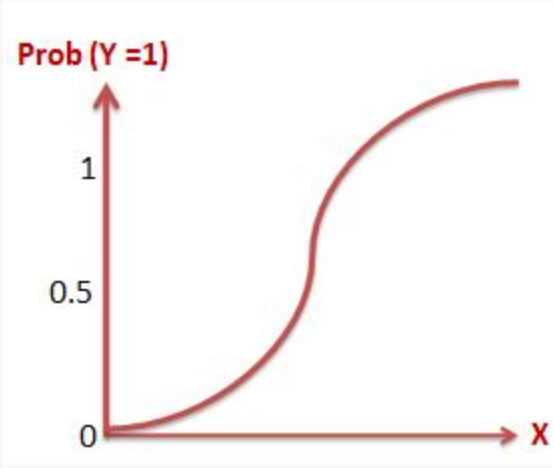
where logit(p) = loge(p / (1-p)

Take exponential both the sides

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + --- + b_k X_k)}}$$

Logistic Regression Equation

p : *the probability of the dependent variable equaling a "success" or "event".*



Logistic Regression Curve

Distribution

Binary logistic regression model assumes binomial distribution of the response with N (number of trials) and p(probability of success). Logistic regression is in the 'binomial family' of GLMs. The dependent variable does not need to be normally distributed.

Example - If you flip a coin twice, what is the probability of getting one or more heads? It's a binomial distribution with N=2 and p=0.5. The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of p.

Interpretation of Logistic Regression Estimates

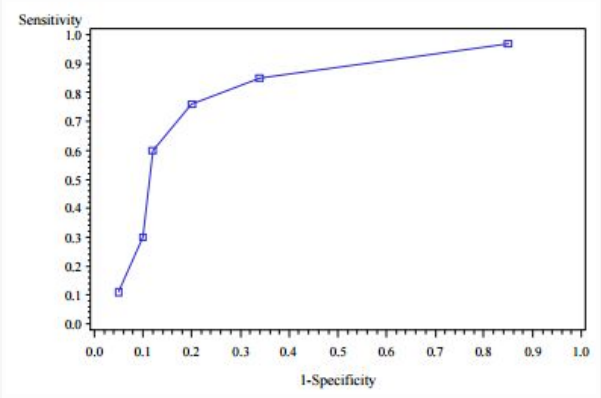
If X increases by one unit, the log-odds of Y increases by k unit, given the other variables in the model are held constant.
In logistic regression, the odds ratio is easier to interpret. That is also called Point estimate. It is exponential value of estimate.

Assumptions of Logistic Regression

1. *The logit transformation of the outcome variable has a linear relationship with the predictor variables.* The one way to check the assumption is to categorize the independent variables. Transform the numeric variables to 10/20 groups and then check whether they have linear or monotonic relationship.
2. *No multicollinearity problem***.** No high correlationship between predictors.
3. *No influential observations (Outliers).*
4. *Large Sample Size* **-** It requires at least 10 events per independent variable.

Model performance:

Area under curve (c statistics) - It plots true positive rate (aka Sensitivity) and false positive rate (aka 1-Specificity). Mathematically, It is calculated using the formula below :



It ranges from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response.

*C = Area under Curve = %concordant + (0.5 * %tied)*
.90-1 = excellent (A)
.80-.90 = good (B)
.70-.80 = fair (C)
.60-.70 = poor (D)
.50-.60 = fail (F)

Classification Table (Confusion Matrix)

Sensitivity (True Positive Rate) - % of events of dependent variable successfully predicted as events.

Sensitivity = TRUE POS / (TRUE POS + FALSE NEG)

Specificity (True Negative Rate) - % of non-events of dependent variable successfully predicted as non-events.

Specificity = TRUE NEG / (TRUE NEG + FALSE POS)

Correct (Accuracy) = Number of correct prediction (TRUE POS + TRUE NEG) divided by sample size.

Difference between Linear and Logistic Regression

1. Variable Type : Linear regression requires the dependent variable to be continuous i.e. numeric values (no categories or groups).

While Binary logistic regression requires the dependent variable to be binary - two categories only (0/1). Multinominal or ordinary logistic regression can have dependent variable with more than two categories.

2. Algorithm : Linear regression is based on ***least square estimation*** which says regression coefficients should be chosen in such a way that it minimizes the sum of the squared distances of each observed response to its fitted value.

While logistic regression is based on ***Maximum Likelihood Estimation*** which says coefficients should be chosen in such a way that it maximizes the **Probability of Y given X** (likelihood). With ML, the computer uses different "iterations" in which it tries different solutions until it gets the maximum likelihood estimates.

3. Equation :

Multiple Regression Equation :

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k$$

Linear Regression Equation

Y is target or dependent variable, b0 is intercept. x1,x2,x3...xk are predictors or independent variables. b1,b2,b3....bk is coefficients of respective predictors.

Logistic Regression Equation :

P(y=1) = e(b0 + b1x1 + b2x2 +-----bkxk) / (1+e(b0+b1x1+ b2x2+------bkxk))

Which further simplifies to :

P(y=1) = 1 / (1+ exp -(b0+b1x1+ b2x2+------bkxk))

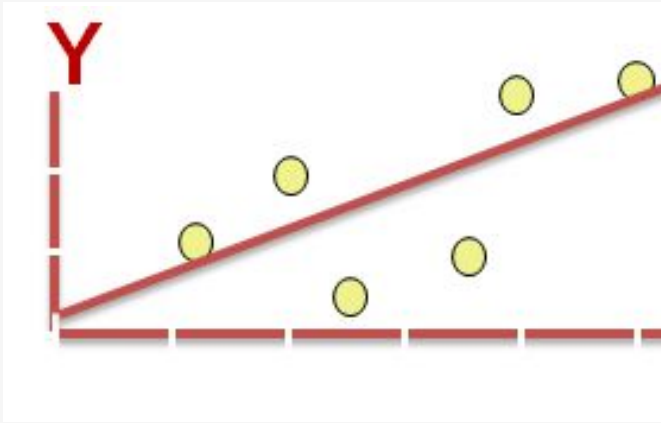
$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + --- + b_k X_k)}}$$

Logistic Regression Equation

The above function is called logistic or sigmoid function.

4. Curve :

Linear Regression : Straight line

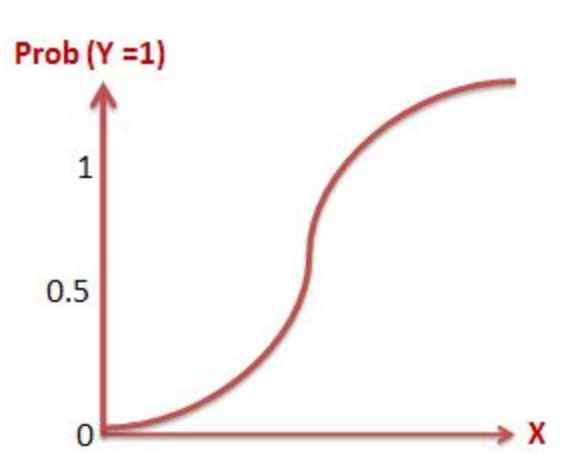


Straight Line : Linear Regression

Linear regression aims at finding the best-fitting straight line which is also called a regression line. In the above figure, the red diagonal line is the

best-fitting straight line and consists of the predicted score on Y for each possible value of X. The distance between the points to the regression line represent the errors.

Logistic Regression : S Curve



Logistic S-shaped curve

Changing the coefficient leads to change in both the direction and the steepness of the logistic function. It means positive slopes result in an S-shaped curve and negative slopes result in a Z-shaped curve.

5. Linear Relationship : Linear regression needs a linear relationship between the dependent and independent variables. While logistic regression does not need a linear relationship between the dependent and independent variables.

6. Normality of Residual : Linear regression requires error term should be normally distributed. While logistic regression does not require error term should be normally distributed.

7. Homoscedasticity : Linear regression assumes that residuals are approximately equal for all predicted dependent variable values. While

Logistic regression does not need residuals to be equal for each level of the predicted dependent variable values.

8. Sample Size : Linear regression requires 5 cases per independent variable in the analysis.While logistic regression needs at least 10 events per independent variable.

9. Purpose : Linear regression is used to estimate the dependent variable incase of a change in independent variables**.** For example, relationship between number of hours studied and your grades.
Whereas logistic regression is used to calculate the probability of an event. For example, an event can be whether customer will attrite or not in next 6 months.

10. Interpretation : Betas or Coefficients of linear regression is interpreted like below -

> *Keeping all other independent variables constant, how much the dependent variable is expected to increase/decrease with an unit increase in the independent variable.*

In logistic regression, we interpret odd ratios -

> *The effect of a one unit of change in X in the predicted odds ratio with the other variables in the model held constant.*

11. Distribution :

Linear regression assumes normal or gaussian distribution of dependent variable. Whereas, **Logistic regression** assumes binomial distribution of dependent variable.