Logistic Regression

The simple and multiple linear regression model is appropriate for relating a quantitative dependent variable (Y) to one or more quantitative independent variables (X). In many regression applications the dependent variable may be qualitative and can assume only discrete values.

Logistic regression is the form of regression that allows the prediction of discrete dependent variables by a mix of continuous and discrete independent variables. The goal of logistic regression is to find the best fitting model to describe the effects of explanatory variables on the qualitative dependent variable.

Logistic regression addresses the same question as multiple regression but with no distributional assumptions on the independent variables (the independent variables do not have to be normally distributed, linearly related or have equal variance in each group) is needed here. Larger samples are needed in logistic regression than for simple or multiple linear regression.

Types of logistic regression:

- **Binary Logistic Regression:** It is used when the dependent variable is dichotomous.
- **Multinomial Logistic Regression**: It is used when the dependent variable has more than two categories.

When and Why Binary Logistic Regression:

We use binary logistic regression in the following situations:

- When the dependent variable is non-parametric.
- When the variance of dependent and independent variable is not equal.
- When the dependent variable has only two levels.
- If we don’t have linearity between dependent and independent variables.

Uses of Binary Logistic Regression:

In binary logistic regression, the dependent variable is binary which assumes two discrete values.

For example, a bank might like to develop an estimated regression equation for predicting whether a person will be approved for a credit card. The dependent variable can be coded as $y = 1$ if the bank approves the request for a credit card and $y = 0$ if the bank rejects the request for a credit card. Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

If a doctor wants to accurately diagnose a possibly cancerous tumor, he can use binary logistic regression to determine whether the tumor is more likely to be benign or malignant.

Logistic Regression Equation:

Suppose we want to assess the relationship between a binary dependent variable y and p independent variables $x_1, x_2, \ldots, x_p$. In logistic regression, statistical theory as well as practice has shown that the relationship between $E(y)$ and $x_1, x_2, \ldots, x_p$ is better described by the following nonlinear equation.

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \cdots (1)$$

The above equation is known as **logistic regression equation.**

If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in Equation (1) provides the probability that $y = 1$ given a particular set of values for the independent variables $x_1, x_2, \ldots, x_p$. Because of the interpretation of $E(y)$ as a probability, the logistic regression equation is often written as follows.

$$E(y) = P(y = 1|x_1, x_2, \ldots, x_p)$$

Estimated Logistic Regression Equation:

The nonlinear form of the logistic regression equation makes the method of computing estimates more complex and beyond the scope of this lesson. We will use computer software to provide the estimates.

The estimated logistic regression equation is

$$\hat{y} = estimate\ of\ P(y = 1|x_1, x_2, \ldots, x_p) = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}$$

where, $b_0, b_1, b_2, \ldots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and

$\hat{y}$ is the estimated value of the dependent variable.

Interpretation of Logistic Regression Equation:

In logistic regression, it is difficult to interpret the relation between the independent variables and the probability that y = 1 directly because the logistic regression equation is nonlinear.

However, statisticians have shown that the relationship can be interpreted indirectly using a concept called the odds ratio. The odds in favor of an event occurring is defined as the probability the event will occur divided by the probability the event will not occur. In logistic regression the event of interest is always y = 1.

Odds Ratio:

Given a particular set of values for the independent variables, the odds in favor of y = 1 can be calculated as follows:

$$odds = \frac{P(y = 1|x_1, x_2, \ldots, x_p)}{P(y = 0|x_1, x_2, \ldots, x_p)} = \frac{P(y = 1|x_1, x_2, \ldots, x_p)}{1 - P(y = 1|x_1, x_2, \ldots, x_p)}$$

The odds ratio measures the impact on the odds of a one-unit increase in only one of the independent variables. The odds ratio is the odds that y = 1 given that one of the independent variables has been increased by one unit ($odds_1$) divided by the odds that y = 1 given no change in the values for the independent variables ($odds_0$).

$$Odds\ Ratio = \frac{odds_1}{odds_0}$$

A unique relationship exists between the odds ratio for a variable and its corresponding regression coefficient. For each independent variable in a logistic regression equation it can be shown that

$$Odds\ Ratio = e^{\beta_j}\ (j = 1,2,\cdots,p)$$

In general, the odds ratio enables us to compare the odds for two different events. If the value of the odds ratio is 1, the odds for both events are the same. Thus, if the independent variable we are considering has a positive impact on the probability of the event occurring, the corresponding odds ratio will be greater than 1.

Logit:

An interesting relationship can be observed between the odds in favor of y = 1 and the exponent for *e* in the logistic regression equation. It can be shown that

$$\ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

This equation shows that the natural logarithm of the odds in favor of y = 1 is a linear function of the independent variables. This linear function is called the logit. We will use the notation $g(x_1, x_2, \ldots, x_p)$ to denote the logit.

Logit: $g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

Once we estimate the parameters in the logistic regression equation, we can compute an estimate of the logit. Using $\hat{g}(x_1, x_2, \ldots, x_p)$ to denote the estimated logit, we obtain

Estimated Logit: $\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$

As logistic regression gives the formula to predict a logit transformation of probability (p) of presence of character of interest, so the model is,

$$logit(p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

In logistic regression the dependent variable is in fact a logit, which is log of odds,

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

So the required probability is

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}}$$

Merits and Demerits of Logistic Regression:

Merits

- It doesn't require too many computational resources.
- It is easily interpretable.
- It is incredibly easy and quick to implement.

Demerits

- It is not a useful tool unless we have already identified all the important independent variables.
- It can be outperformed by the more complex situations.
- It doesn't perform well with independent variables that are not correlated to dependent variable.

Example

With the logistic regression equation, we can model the probability of a manual transmission in a vehicle based on its weight and engine horsepower.

We will use logistic regression to compute the odds of have a manual transmission rather than an automatic transmission, given the weight and gross horsepower of the automobiles.

Here the **dependent variable** is Transmission Type of the automobile model [0=Automatic, 1=manual]

Independent Variables: Weight of automobile engines (in pounds) and horsepower of the automobile engines.

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Output from R

glm(formula = am ~ hp + wt, family = binomial(link = "logit"), data = mtcars)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2537	-0.1568	-0.0168	0.1543	1.3449

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	18.86630	7.44356	2.535	0.01126 *
hp	0.03626	0.01773	2.044	0.04091 *
wt	-8.08348	3.06868	-2.634	0.00843 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 43.230 on 31 degrees of freedom

Residual deviance: 10.059 on 29 degrees of freedom

Output for Odds Ratio

(Intercept)	hp	wt
1.561455e+08	1.036921e+00	3.085967e-04

Interpretation:

The estimated logistic regression equation is

$$\hat{y} = \frac{e^{18.87+0.036x_1-8.083x_2}}{1+e^{18.87+0.036x_1-8.083x_2}}$$

We can interpret logistic regression in terms of odds ratio.

- The odds of having a manual transmission, rather than an automatic transmission, increase by 1.04 for every unit increase in gross horse power.
- The odds of having a manual transmission, rather than an automatic transmission, are by 0.0003 for every 1000 pounds increase in car weight.

We can also calculate the odds ratio using the following formula:

$$Odds\ Ratio = e^{\beta_j}\ (j = 1,2,\cdots,p)$$

As $b_1 = 0.03626$,

$$Odds\ Ratio = e^{0.03626} \cong 1.0367$$

As $b_2 = $ -8.08348,

$$Odds\ Ratio = e^{-8.08348} \cong 0.0003$$

Practice Problem:

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores) and GPA (grade point average) effect admission into graduate school. The response variable, admit/don't admit, is a binary variable. Suppose we have the following output for this problem.

```
Call:
glm(formula = admit ~ gre + gpa, family = "binomial", data = mydata)

  Deviance Residuals:
     Min      1Q  Median      3Q     Max
  -1.627  -0.866  -0.639   1.149   2.079
Coefficients:
                   Estimate Std. Error  z value   Pr(>|z|)
  (Intercept) -3.98998    1.13995   -3.50     0.00047 ***
   gre          0.00226     0.00109    2.07      0.03847 *
   gpa          0.80404     0.33182    2.42      0.01539 *

 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, determine the estimated logistic regression equation and interpret it.