Introduction to Logistic Regression

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables.

The name logistic regression is used when the dependent variable has only two values, such as

- {0 and 1} or
- {Yes and No} or
- {Spam and No-Spam}

The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as

- {Married, Single, Divored, Widowed}
- {Cancer, No-Cancer, No-Answer}

Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

The Logit and Logistic Transformations:

In multiple regression, a mathematical model of a set of explanatory variables is used to

predict the mean of a continuous dependent
variable.

In logistic regression, a mathematical model
of a set of explanatory variables is used to
predict a logit transformation of the dependent
variable.

Labels: 0 and 1
==========
Suppose the numerical values of 0 and 1 are
assigned to the two outcomes of a binary variable.

Often,
    the 0 represents a negative response and
    the 1 represents a positive response.
    The mean of this variable will be the
    proportion of positive responses.

p and (1-p): the logarithm of the odds
==========
If p is the proportion of observations with
an outcome of 1, then (1-p) is the probability
of a outcome of 0.

N = All positive and negative responses
1 => positive_responses = A
0 => negative_responses = B
N = A + B
p = A / N
1 - p = (B / N)

The ratio p/(1-p) is called the odds and the

logit is the logarithm of the odds, or just
log odds.

Logit Transformation

Mathematically, the logit transformation is written

Let log be logarithm notation

$$L = \text{logit}(p) = \log\left(p / (1-p)\right)$$

In mathematics, the logarithm is the inverse function
to exponentiation.

$$\log_b(x) = y$$

$$b^y = x$$

$$\log_2(64) = 6 \iff 2^6 = 64$$

$$\log_2(32) = 5 \iff 2^5 = 32$$

$$\log_5(25) = 2 \iff 5^2 = 25$$

The following table shows the logit for various values of p.

P	Logit(P)	
0.001	-6.907	
0.01	-4.595	
0.05	-2.944	
0.10	-2.197	
0.20	-1.386	
0.30	-0.847	
0.40	-0.405	
0.50	0.000	<======
0.60	0.405	
0.70	0.847	
0.80	1.386	
0.90	2.197	
0.95	2.944	
0.99	4.595	
0.999	6.907	

Range of p
==========
Range of p:

p ranges between zero (0.00) and one (1.00)

Range of logit:

logit ranges between minus and plus infinity oo

the zero logit occurs when p is 0.50

The logistic transformation

The logistic transformation is the inverse of the logit transformation.

It is written

$$p = \text{logistic}(L) = (e^L) / (1 + (e^L))$$

The Log Odds Ratio Transformation

The difference between two log odds can be used to compare two proportions, such as that of males versus females. Mathematically, this difference is written

$$\begin{aligned} L1 - L2 &= \text{logit}(p1) - \text{logit}(p2) \\ &= \log(p1 / (1-p1)) - \log(p2 / (1-p2)) \\ &= \log\left((p1/(1-p1)) / (p2/(1-p2)) \right) \\ &= \log\left(p1(1-p2) / p2(1-p1) \right) \end{aligned}$$

The Logistic Regression and Logit Models

In logistic regression, a categorical dependent variable Y having G (usually G = 2) unique values is regressed on a set of p independent variables

X1 X2 Xp , ,..., .

For example, Y may be presence or absence of a disease, condition after surgery, or marital status. Since the names of these partitions are arbitrary, we often refer to them by consecutive numbers. That is, in the discussion below, Y will take on the values 1, 2, … G. In fact,

Let X = (X1, X2, ..., Xp)

Logistic Model:
==========

Let us try to understand logistic regression by considering a logistic model with given parameters, then seeing how the coefficients can be estimated from data. Consider a model with two predictors, x_1 and x_2, and one binary response variable Y, which we denote:

p = P(Y=1)

L = log (p / (1-p)) = B0 + B1 x_1 + B2 x_2
b

b = 2, 10, ...

(p / (1-p)) = b ^ (B0 + B1 x_1 + B2 x_2)

By simple algebraic manipulation,

the probability that Y=1 is

$$p = [\, b \wedge (B0 + B1\, x\_1 + B2\, x\_2) \,] \,/\, [\, b \wedge (B0 + B1\, x\_1 + B2\, x\_2) + 1]$$

$$= 1 \,/\, [1 + b \wedge -(B0 + B1\, x\_1 + B2\, x\_2)]$$

The above formula shows that once B_{i}
(B0, B1, B2) are fixed, we can easily
compute either the log-odds that (Y = 1)
for a given observation, or the probability
that (Y = 1) for a given observation.

The main use-case of a logistic model:

The main use-case of a logistic model is
to be given an observation (x_1, x_2), and
estimate the probability p that Y=1.

In most applications, the base b of the
logarithm is usually taken to be e. However
in some cases it can be easier to communicate
results by working in base 2, or base 10.

We consider an example with b=10, and coefficients
B_0 = -3
B_1 = 1
B_2 = 2

To be concrete, the model is

$$L = \log\left(\frac{p}{1-p}\right) = -3 + X_1 + 2 X_2$$
10

where p is the probability of the event that Y=1.

Estimation:
==========
In order to estimate the parameters {B_0, B_1, B_2} from data, one must do logistic regression.

Example:

Probability of passing an exam versus hours of study

To answer the following question:

A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used.

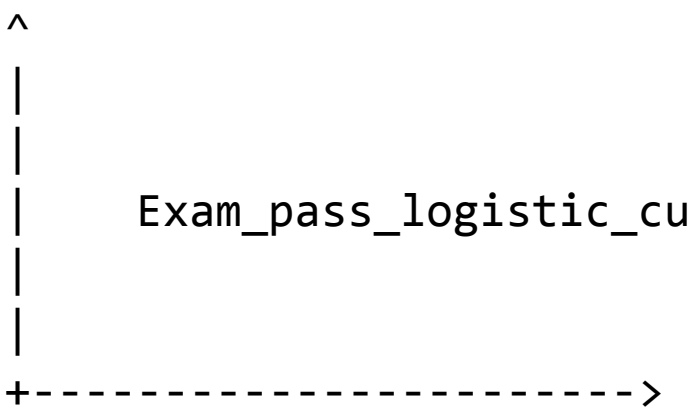
The table shows the number of hours each student spent

studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75
1.75	2.00	2.25	2.50	2.75	3.00	3.25
3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0
1	0	1	0	1	0	1
1	1	1	1	1		

Hours	Pass
0.50	0
0.75	0
1.0	0
1.25	0
1.50	0
1.75	0
1.75	0
2.00	0
2.25	1
2.50	0
2.75	1
3.00	0
3.25	1
3.50	0
4.00	1
4.25	1
4.75	1
5.00	1
5.50	1

The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.



probability >= 50% => Y = 1
probability < 50% => Y = 0

logit (p) = log (p / (1-p))   for 0 < p < 1