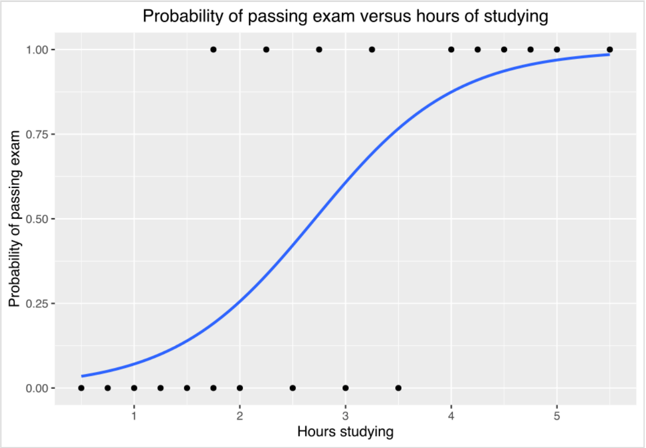
Logistic regression

In statistics, a **logistic model** (or **logit model**) is a statistical model that models the log-odds of an event as a linear combination of one or more independent variables. In regression analysis, **logistic regression**[1] (or **logit regression**) estimates the parameters of a logistic model (the coefficients in the linear or non linear combinations). In binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling;[2] the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from ***logistic unit***, hence the alternative names. See § Background and § Definition for formal mathematics, and § Example for a worked example.



Example graph of a logistic regression curve fitted to data. The curve shows the estimated probability of passing an exam (binary dependent variable) versus hours studying (scalar independent variable). See § Example for worked details.

Binary variables are widely used in statistics to model the probability of a certain class or event taking place, such as the probability of a team winning, of a patient being healthy, etc. (see § Applications), and the logistic model has been the most commonly used model for binary regression since about 1970.[3] Binary variables can be generalized to categorical variables when there are more than two possible values (e.g. whether an image is of a cat, dog, lion, etc.), and the binary logistic regression generalized to multinomial logistic regression. If the multiple categories are ordered, one can use the ordinal logistic regression (for example the proportional odds ordinal logistic model[4]). See § Extensions for further extensions. The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

Analogous linear models for binary variables with a different sigmoid function instead of the logistic function (to convert the linear combination to a probability) can also be used, most notably the probit model; see § Alternatives. The defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio. More abstractly, the logistic function is the natural parameter for the Bernoulli distribution, and in this sense is the "simplest" way to convert a real number to a probability.

The parameters of a logistic regression are most commonly estimated by maximum-likelihood estimation (MLE). This does not have a closed-form expression, unlike linear least squares; see § Model fitting. Logistic regression by MLE plays a similarly basic role for binary or categorical responses as linear regression by ordinary least squares (OLS) plays for scalar responses: it is a simple, well-analyzed baseline model; see § Comparison with linear regression for discussion. The logistic regression as a general statistical model was originally developed and popularized primarily by Joseph Berkson,[5] beginning in Berkson (1944), where he coined "logit"; see § History.

Applications

General

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd *et al.* using logistic regression.[6] Many other medical scales used to assess severity of a patient have been developed using logistic regression.[7][8][9][10] Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.).[11][12] Another example might be to predict whether a Nepalese voter will vote Nepali Congress or Communist Party of Nepal or for any other party, based on age, income, sex, race, state of residence, votes in previous elections, etc.[13] The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product.[14][15] It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc.[16] In economics, it can be used to predict the likelihood of a person ending up in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing. Disaster planners and engineers rely on these models to predict decisions taken by householders or building occupants in small-scale and large-scales evacuations, such as building fires, wildfires, hurricanes among others.[17][18][19] These models help in the development of reliable disaster managing plans and safer design for the built environment.

Supervised machine learning

Logistic regression is a supervised machine learning algorithm widely used for binary classification tasks, such as identifying whether an email is spam or not and diagnosing diseases by assessing the presence or absence of specific conditions based on patient test results. This approach utilizes the logistic (or sigmoid) function to transform a linear combination of input features into a probability value ranging between 0 and 1. This probability indicates the likelihood that a given input corresponds to one of two predefined categories. The essential mechanism of logistic regression is grounded in the logistic function's ability to model the probability of binary outcomes accurately. With its distinctive S-shaped curve, the logistic function effectively maps any real-valued number to a value within the 0 to 1 interval. This feature renders it particularly suitable for binary classification tasks, such as sorting emails into "spam" or "not spam". By calculating the probability that the dependent variable will be categorized into a specific group, logistic regression provides a probabilistic framework that supports informed decision-making.[20]

Example

Problem

As a simple example, we can use a logistic regression with one explanatory variable and two categories 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 (x_k)	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 (y_k)	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

We wish to fit a logistic function to the data consisting of the hours studied (x_k) and the outcome of the test (y_k =1 for pass, 0 for fail). The data points are indexed by the subscript k which runs from $k = 1$ to $k = K = 20$. The x variable is called the "explanatory variable", and the y variable is called the "categorical variable" consisting of two categories: "pass" or "fail" corresponding to the categorical values 1 and 0 respectively.

Model

The logistic function is of the form:

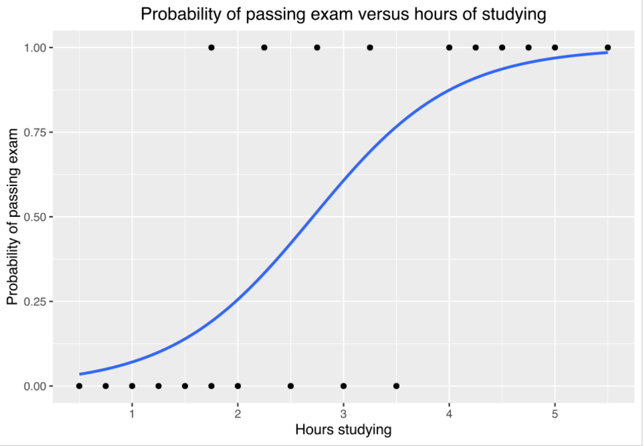
$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

where μ is a location parameter (the midpoint of the curve, where $p(\mu) = 1/2$) and s is a scale parameter. This expression may be rewritten as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where $\beta_0 = -\mu/s$ and is known as the intercept (it is the *vertical* intercept or y-intercept of the line $y = \beta_0 + \beta_1 x$), and $\beta_1 = 1/s$ (inverse scale parameter or rate parameter): these are the y-intercept and slope of the log-odds as a function of x. Conversely, $\mu = -\beta_0/\beta_1$ and $s = 1/\beta_1$.

Note that this model is actually an oversimplification, as it implies that every student will pass if they study indefinitely (limit = 1).



Graph of a logistic regression curve fitted to the (x_m,y_m) data. The curve shows the probability of passing an exam versus hours studying.

Fit

The usual measure of goodness of fit for a logistic regression uses logistic loss (or log loss), the negative log-likelihood. For a given x_k and y_k, write $p_k = p(x_k)$. The p_k are the probabilities that the corresponding y_k will equal one and $1 - p_k$ are the probabilities that they will be zero (see Bernoulli distribution). We wish to find the values of β_0 and β_1 which give the "best fit" to the data. In the case of linear regression, the sum of the squared deviations of the fit from the data points (y_k), the squared error loss, is taken as a measure of the goodness of fit, and the best fit is obtained when that function is *minimized*.

The log loss for the k-th point ℓ_k is:

$$\ell_k = \begin{cases} -\ln p_k & \text{if } y_k = 1, \\ -\ln(1 - p_k) & \text{if } y_k = 0. \end{cases}$$

The log loss can be interpreted as the "surprisal" of the actual outcome y_k relative to the prediction p_k, and is a measure of information content. Log loss is always greater than or equal to 0, equals 0 only in case of a perfect prediction (i.e., when $p_k = 1$ and $y_k = 1$, or $p_k = 0$ and $y_k = 0$), and approaches infinity as the prediction gets worse (i.e., when $y_k = 1$ and $p_k \to 0$ or $y_k = 0$ and $p_k \to 1$), meaning the actual outcome is "more surprising". Since the value of the logistic function is always strictly between zero and one, the log loss is always greater than zero and less than infinity. Unlike in a linear regression, where the model can have zero loss at a point by passing through a data point (and zero loss overall if all points are on a line), in a logistic regression it is not possible to have zero loss at any points, since y_k is either 0 or 1, but $0 < p_k < 1$.

These can be combined into a single expression:

$$\ell_k = -y_k \ln p_k - (1 - y_k) \ln(1 - p_k).$$

This expression is more formally known as the cross-entropy of the predicted distribution $\big(p_k, (1 - p_k)\big)$ from the actual distribution $\big(y_k, (1 - y_k)\big)$, as probability distributions on the two-element space of (pass, fail).

The sum of these, the total loss, is the overall negative log-likelihood $-\ell$, and the best fit is obtained for those choices of β_0 and β_1 for which $-\ell$ is *minimized.*

Alternatively, instead of *minimizing* the loss, one can *maximize* its inverse, the (positive) log-likelihood:

$$\ell = \sum_{k:y_k=1} \ln(p_k) + \sum_{k:y_k=0} \ln(1 - p_k) = \sum_{k=1}^{K} \left(y_k \ln(p_k) + (1 - y_k) \ln(1 - p_k) \right)$$

or equivalently maximize the likelihood function itself, which is the probability that the given data set is produced by a particular logistic function:

$$L=\prod_{k:y_k=1}p_k\prod_{k:y_k=0}(1-p_k)$$

This method is known as maximum likelihood estimation.

Parameter estimation

Since ℓ is nonlinear in β_0 and β_1, determining their optimum values will require numerical methods. One method of maximizing ℓ is to require the derivatives of ℓ with respect to β_0 and β_1 to be zero:

$$0=\frac{\partial\ell}{\partial\beta_0}=\sum_{k=1}^{K}(y_k-p_k)$$

$$0=\frac{\partial\ell}{\partial\beta_1}=\sum_{k=1}^{K}(y_k-p_k)x_k$$

and the maximization procedure can be accomplished by solving the above two equations for β_0 and β_1, which, again, will generally require the use of numerical methods.

The values of β_0 and β_1 which maximize ℓ and L using the above data are found to be:

$$\beta_0\approx-4.1$$
$$\beta_1\approx1.5$$

which yields a value for μ and s of:

$$\mu=-\beta_0/\beta_1\approx2.7$$
$$s=1/\beta_1\approx0.67$$

Predictions

The β_0 and β_1 coefficients may be entered into the logistic regression equation to estimate the probability of passing the exam.

For example, for a student who studies 2 hours, entering the value $x=2$ into the equation gives the estimated probability of passing the exam of 0.25:

$$t=\beta_0+2\beta_1\approx-4.1+2\cdot1.5=-1.1$$

$$p=\frac{1}{1+e^{-t}}\approx0.25=\text{Probability of passing exam}$$

Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$t=\beta_0+4\beta_1\approx-4.1+4\cdot1.5=1.9$$

$$p=\frac{1}{1+e^{-t}}\approx0.87=\text{Probability of passing exam}$$

This table shows the estimated probability of passing the exam for several values of hours studying.

Hours of study (x)	Passing exam		
	Log-odds (t)	Odds (e^t)	Probability (p)
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
$\mu \approx 2.7$	0	1	$\frac{1}{2}$ = 0.50
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

Model evaluation

The logistic regression analysis gives the following output.

	Coefficient	Std. Error	z-value	p-value (Wald)
Intercept (β_0)	−4.1	1.8	−2.3	0.021
Hours (β_1)	1.5	0.9	1.7	0.017

By the Wald test, the output indicates that hours studying is significantly associated with the probability of passing the exam ($p = 0.017$). Rather than the Wald method, the recommended method[21] to calculate the p-value for logistic regression is the likelihood-ratio test (LRT), which for these data give $p \approx 0.00064$ (see § Deviance and likelihood ratio tests below).

Generalizations

This simple model is an example of binary logistic regression, and has one explanatory variable and a binary categorical variable which can assume one of two categorical values. Multinomial logistic regression is the generalization of binary logistic regression to include any number of explanatory variables and any number of categories.

Background

Definition of the logistic function

An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is a sigmoid function, which takes any real input t, and outputs a value between zero and one.[2] For the logit, this is interpreted as taking input log-odds and having output probability. The *standard* logistic function $\sigma : \mathbb{R} \rightarrow (0,1)$ is defined as follows:

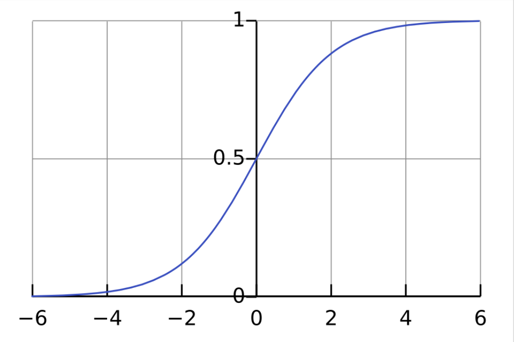


Figure 1. The standard logistic function $\sigma(t)$; $\sigma(t) \in (0,1)$ for all t.

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

A graph of the logistic function on the t-interval (−6,6) is shown in Figure 1.

Let us assume that t is a linear function of a single explanatory variable x (the case where t is a *linear combination* of multiple explanatory variables is treated similarly). We can then express t as follows:

$$t = \beta_0 + \beta_1 x$$

And the general logistic function $p : \mathbb{R} \rightarrow (0,1)$ can now be written as:

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

In the logistic model, $p(x)$ is interpreted as the probability of the dependent variable Y equaling a success/case rather than a failure/non-case. It is clear that the response variables Y_i are not identically distributed: $P(Y_i = 1 \mid X)$ differs from one data point X_i to another, though they are independent given design matrix X and shared parameters β.[11]

Definition of the inverse of the logistic function

We can now define the logit (log odds) function as the inverse $g = \sigma^{-1}$ of the standard logistic function. It is easy to see that it satisfies:

$$g(p(x)) = \sigma^{-1}(p(x)) = \operatorname{logit} p(x) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x,$$

and equivalently, after exponentiating both sides we have the odds:

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}.$$

Interpretation of these terms

In the above equations, the terms are as follows:

- g is the logit function. The equation for $g(p(x))$ illustrates that the logit (i.e., log-odds or natural logarithm of the odds) is equivalent to the linear regression expression.
- $\ln$ denotes the natural logarithm.
- $p(x)$ is the probability that the dependent variable equals a case, given some linear combination of the predictors. The formula for $p(x)$ illustrates that the probability of the dependent variable equaling a case is equal to the value of the logistic function of the linear regression expression. This is important in that it shows that the value of the linear regression expression can vary from negative to positive infinity and yet, after transformation, the resulting expression for the probability $p(x)$ ranges between 0 and 1.
- β_0 is the intercept from the linear regression equation (the value of the criterion when the predictor is equal to zero).
- $\beta_1 x$ is the regression coefficient multiplied by some value of the predictor.
- base e denotes the exponential function.

Definition of the odds

The odds of the dependent variable equaling a case (given some linear combination x of the predictors) is equivalent to the exponential function of the linear regression expression. This illustrates how the logit serves as a link function between the probability and the linear regression expression. Given that the logit ranges between negative and positive infinity, it provides an adequate criterion upon which to conduct linear regression and the logit is easily converted back into the odds.[2]

So we define odds of the dependent variable equaling a case (given some linear combination x of the predictors) as follows:

$$\text{odds} = e^{\beta_0 + \beta_1 x}.$$

The odds ratio

For a continuous independent variable the odds ratio can be defined as:

$$\mathrm{OR} = \frac{\operatorname{odds}(x+1)}{\operatorname{odds}(x)} = \frac{\left(\frac{p(x+1)}{1-p(x+1)}\right)}{\left(\frac{p(x)}{1-p(x)}\right)} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

This exponential relationship provides an interpretation for β_1: The odds multiply by e^{β_1} for every 1-unit increase in x.[22]

For a binary independent variable the odds ratio is defined as $\frac{ad}{bc}$ where a, b, c and d are cells in a 2×2 contingency table.[23]

Multiple explanatory variables

If there are multiple explanatory variables, the above expression $\beta_0 + \beta_1 x$ can be revised to $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$. Then when this is used in the equation relating the log odds of a success to the values of the predictors, the linear regression will be a multiple regression with m explanators; the parameters β_i for all $i = 0, 1, 2, \dots, m$ are all estimated.

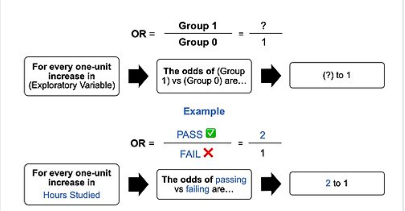
Again, the more traditional equations are:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

and

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

where usually $b = e$.


OR = Group 1 / Group 0 = ? / 1
For every one-unit increase in (Exploratory Variable)
The odds of (Group 1) vs (Group 0) are...
(?) to 1
Example
OR = PASS ✓ / FAIL ✗ = 2 / 1
For every one-unit increase in Hours Studied
The odds of passing vs failing are...
2 to 1

The image represents an outline of what an odds ratio looks like in writing, through a template in addition to the test score example in the "Example" section of the contents. In simple terms, if we hypothetically get an odds ratio of 2 to 1, we can say... "For every one-unit increase in hours studied, the odds of passing (group 1) or failing (group 0) are (expectedly) 2 to 1 (Denis, 2019).

Definition

A dataset contains N points. Each point i consists of a set of m input variables $x_{1,i} \dots x_{m,i}$ (also called independent variables, explanatory variables, predictor variables, features, or attributes), and a binary outcome variable Y_i (also known as a dependent variable, response variable, output variable, or class), i.e. it can assume only the two possible values 0 (often meaning "no" or "failure") or 1 (often meaning "yes" or "success"). The goal of logistic regression is to use the dataset to create a predictive model of the outcome variable.

As in linear regression, the outcome variables Y_i are assumed to depend on the explanatory variables $x_{1,i} \dots x_{m,i}$.

Explanatory variables

The explanatory variables may be of any type: real-valued, binary, categorical, etc. The main distinction is between continuous variables and discrete variables.

(Discrete variables referring to more than two possible choices are typically coded using dummy variables (or indicator variables), that is, separate explanatory variables taking the value 0 or 1 are created for each possible value of the discrete variable, with a 1 meaning "variable does have the given value" and a 0 meaning "variable does not have that value".)

Outcome variables

Formally, the outcomes Y_i are described as being Bernoulli-distributed data, where each outcome is determined by an unobserved probability p_i that is specific to the outcome at hand, but related to the explanatory variables. This can be expressed in any of the following equivalent forms:

$$\begin{aligned}
Y_i \mid x_{1,i}, \ldots, x_{m,i} \ &\sim \operatorname{Bernoulli}(p_i) \\
\mathbb{E}[Y_i \mid x_{1,i}, \ldots, x_{m,i}] &= p_i \\
\Pr(Y_i = y \mid x_{1,i}, \ldots, x_{m,i}) &= \begin{cases} p_i & \text{if } y = 1 \\ 1 - p_i & \text{if } y = 0 \end{cases} \\
\Pr(Y_i = y \mid x_{1,i}, \ldots, x_{m,i}) &= p_i^y (1 - p_i)^{(1-y)}
\end{aligned}$$

The meanings of these four lines are:

1. The first line expresses the probability distribution of each Y_i : conditioned on the explanatory variables, it follows a Bernoulli distribution with parameters p_i, the probability of the outcome of 1 for trial i. As noted above, each separate trial has its own

probability of success, just as each trial has its own explanatory variables. The probability of success p_i is not observed, only the outcome of an individual Bernoulli trial using that probability.

2. The second line expresses the fact that the expected value of each Y_i is equal to the probability of success p_i, which is a general property of the Bernoulli distribution. In other words, if we run a large number of Bernoulli trials using the same probability of success p_i, then take the average of all the 1 and 0 outcomes, then the result would be close to p_i. This is because doing an average this way simply computes the proportion of successes seen, which we expect to converge to the underlying probability of success.
3. The third line writes out the probability mass function of the Bernoulli distribution, specifying the probability of seeing each of the two possible outcomes.
4. The fourth line is another way of writing the probability mass function, which avoids having to write separate cases and is more convenient for certain types of calculations. This relies on the fact that Y_i can take only the value 0 or 1. In each case, one of the exponents will be 1, "choosing" the value under it, while the other is 0, "canceling out" the value under it. Hence, the outcome is either p_i or 1 – p_i, as in the previous line.

Linear predictor function

The basic idea of logistic regression is to use the mechanism already developed for linear regression by modeling the probability p_i using a linear predictor function, i.e. a linear combination of the explanatory variables and a set of regression coefficients that are specific to the model at hand but the same for all trials. The linear predictor function $f(i)$ for a particular data point i is written as:

$$f(i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i},$$

where $\beta_0, \ldots, \beta_m$ are regression coefficients indicating the relative effect of a particular explanatory variable on the outcome.

The model is usually put into a more compact form as follows:

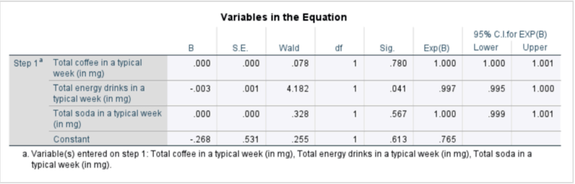
- The regression coefficients β_0, β_1, ..., β_m are grouped into a single vector $\boldsymbol{\beta}$ of size $m + 1$.
- For each data point i, an additional explanatory pseudo-variable $x_{0,i}$ is added, with a fixed value of 1, corresponding to the intercept coefficient β_0.
- The resulting explanatory variables $x_{0,i}$, $x_{1,i}$, ..., $x_{m,i}$ are then grouped into a single vector $\boldsymbol{X_i}$ of size $m + 1$.

This makes it possible to write the linear predictor function as follows:

$$f(i) = \boldsymbol{\beta} \cdot \mathbf{X}_i,$$

using the notation for a dot product between two vectors.

Many explanatory variables, two categories



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Total coffee in a typical week (in mg)	.000	.000	.078	1	.780	1.000	1.000	1.001
	Total energy drinks in a typical week (in mg)	-.003	.001	4.182	1	.041	.997	.995	1.000
	Total soda in a typical week (in mg)	.000	.000	.328	1	.567	1.000	.999	1.001
	Constant	-.268	.531	.255	1	.613	.765		

a. Variable(s) entered on step 1: Total coffee in a typical week (in mg), Total energy drinks in a typical week (in mg), Total soda in a typical week (in mg).

This is an example of an SPSS output for a logistic regression model using three explanatory variables (coffee use per week, energy drink use per week, and soda use per week) and two categories (male and female).

The above example of binary logistic regression on one explanatory variable can be generalized to binary logistic regression on any number of explanatory variables x_1, x_2,... and any number of categorical values $y = 0, 1, 2, \dots$.

To begin with, we may consider a logistic model with M explanatory variables, x_1, x_2 ... x_M and, as in the example above, two categorical values (y = 0 and 1). For the simple binary logistic regression model, we assumed a linear relationship between the predictor variable and the log-odds (also called logit) of the event that $y = 1$. This linear relationship may be extended to the case of M explanatory variables:

$$t = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M$$

where t is the log-odds and β_i are parameters of the model. An additional generalization has been introduced in which the base of the model (b) is not restricted to Euler's number e. 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.

For a more compact notation, we will specify the explanatory variables and the β coefficients as $(M+1)$-dimensional vectors:

$$\boldsymbol{x} = \{x_0, x_1, x_2, \dots, x_M\}$$

$$\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2, \dots, \beta_M\}$$

with an added explanatory variable x_0 =1. The logit may now be written as:

$$t = \sum_{m=0}^{M} \beta_m x_m = \boldsymbol{\beta} \cdot x$$

Solving for the probability p that $y = 1$ yields:

$$p(\boldsymbol{x}) = \frac{b^{\boldsymbol{\beta}\cdot\boldsymbol{x}}}{1 + b^{\boldsymbol{\beta}\cdot\boldsymbol{x}}} = \frac{1}{1 + b^{-\boldsymbol{\beta}\cdot\boldsymbol{x}}} = S_b(t),$$

where S_b is the sigmoid function with base b. The above formula shows that once the β_m 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 is to be given an observation $\boldsymbol{x}$, and estimate the probability $p(\boldsymbol{x})$ that $y = 1$. The optimum beta coefficients may again be found by maximizing the log-likelihood. For K measurements, defining $\boldsymbol{x}_k$ as the explanatory vector of the k-th measurement, and y_k as the categorical outcome of that measurement, the log likelihood may be written in a form very similar to the simple $M = 1$ case above:

$$\ell = \sum_{k=1}^{K} y_k \log_b(p(\boldsymbol{x_k})) + \sum_{k=1}^{K}(1 - y_k)\log_b(1 - p(\boldsymbol{x_k}))$$

As in the simple example above, finding the optimum β parameters will require numerical methods. One useful technique is to equate the derivatives of the log likelihood with respect to each of the β parameters to zero yielding a set of equations which will hold at the maximum of the log likelihood:

$$\frac{\partial \ell}{\partial \beta_m} = 0 = \sum_{k=1}^{K} y_k x_{mk} - \sum_{k=1}^{K} p(\boldsymbol{x}_k) x_{mk}$$

where x_{mk} is the value of the x_m explanatory variable from the *k-th* measurement.

Consider an example with $M = 2$ explanatory variables, $b = 10$, and coefficients $\beta_0 = -3$, $\beta_1 = 1$, and $\beta_2 = 2$ which have been determined by the above method. To be concrete, the model is:

$$t = \log_{10} \frac{p}{1-p} = -3 + x_1 + 2x_2$$
$$p = \frac{b^{\boldsymbol{\beta}\cdot\boldsymbol{x}}}{1 + b^{\boldsymbol{\beta}\cdot x}} = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}},$$

where p is the probability of the event that $y = 1$. This can be interpreted as follows:

- $\beta_0 = -3$ is the y-intercept. It is the log-odds of the event that $y = 1$, when the predictors $x_1 = x_2 = 0$. By exponentiating, we can see that when $x_1 = x_2 = 0$ the odds of the event that $y = 1$ are 1-to-1000, or 10^{-3}. Similarly, the probability of the event that $y = 1$ when $x_1 = x_2 = 0$ can be computed as $1/(1000 + 1) = 1/1001.$
- $\beta_1 = 1$ means that increasing x_1 by 1 increases the log-odds by 1. So if x_1 increases by 1, the odds that $y = 1$ increase by a factor of 10^1. The **probability** of $y = 1$ has also increased, but it has not increased by as much as the odds have increased.
- $\beta_2 = 2$ means that increasing x_2 by 1 increases the log-odds by 2. So if x_2 increases by 1, the odds that $y = 1$ increase by a factor of 10^2. Note how the effect of x_2 on the log-odds is twice as great as the effect of x_1, but the effect on the odds is 10 times greater. But the effect on the **probability** of $y = 1$ is not as much as 10 times greater, it's only the effect on the odds that is 10 times greater.

Multinomial logistic regression: Many explanatory variables and many categories

In the above cases of two categories (binomial logistic regression), the categories were indexed by "0" and "1", and we had two probabilities: The probability that the outcome was in category 1 was given by $p(\boldsymbol{x})$and the probability that the outcome was in category 0 was given by $1 - p(\boldsymbol{x})$. The sum of these probabilities equals 1, which must be true, since "0" and "1" are the only possible categories in this setup.

In general, if we have $M+1$ explanatory variables (including x_o) and $N+1$ categories, we will need $N+1$ separate probabilities, one for each category, indexed by n, which describe the probability that the categorical outcome y will be in category y=n, conditional on the vector of covariates **x**. The sum of these probabilities over all categories must equal 1. Using the mathematically convenient base e, these probabilities are:

$$p_n(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}_n \cdot \boldsymbol{x}}}{1 + \sum_{u=1}^{N} e^{\boldsymbol{\beta}_u \cdot \boldsymbol{x}}} \text{ for } n = 1, 2, \dots, N$$

$$p_0(\boldsymbol{x}) = 1 - \sum_{n=1}^{N} p_n(\boldsymbol{x}) = \frac{1}{1 + \sum_{u=1}^{N} e^{\boldsymbol{\beta}_u \cdot \boldsymbol{x}}}$$

Each of the probabilities except $p_0(\boldsymbol{x})$ will have their own set of regression coefficients $\boldsymbol{\beta}_n$. It can be seen that, as required, the sum of the $p_n(\boldsymbol{x})$ over all categories n is 1. The selection of $p_0(\boldsymbol{x})$ to be defined in terms of the other probabilities is artificial. Any of the probabilities could have been selected to be so defined. This special value of n is termed the "pivot index", and the log-odds (t_n) are expressed in terms of the pivot probability and are again expressed as a linear combination of the explanatory variables:

$$t_n = \ln\left(\frac{p_n(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right) = \boldsymbol{\beta}_n \cdot \boldsymbol{x}$$

Note also that for the simple case of $N = 1$, the two-category case is recovered, with $p(\boldsymbol{x}) = p_1(\boldsymbol{x})$ and $p_0(\boldsymbol{x}) = 1 - p_1(\boldsymbol{x})$.

The log-likelihood that a particular set of K measurements or data points will be generated by the above probabilities can now be calculated. Indexing each measurement by k, let the k-th set of measured explanatory variables be denoted by $\boldsymbol{x}_k$ and their categorical outcomes be denoted by y_k which can be equal to any integer in [0,N]. The log-likelihood is then:

$$\ell = \sum_{k=1}^{K} \sum_{n=0}^{N} \Delta(n, y_k) \ln(p_n(\boldsymbol{x}_k))$$

where $\Delta(n, y_k)$ is an indicator function which equals 1 if $y_k = n$ and zero otherwise. In the case of two explanatory variables, this indicator function was defined as y_k when n = 1 and 1-y_k when n = 0. This was convenient, but not necessary.[24] Again, the optimum beta coefficients may be found by maximizing the log-likelihood function generally using numerical methods. A possible method of solution is to set the derivatives of the log-likelihood with respect to each beta coefficient equal to zero and solve for the beta coefficients:

$$\frac{\partial \ell}{\partial \beta_{nm}} = 0 = \sum_{k=1}^{K} \Delta(n, y_k) x_{mk} - \sum_{k=1}^{K} p_n(\boldsymbol{x}_k) x_{mk}$$

where β_{nm} is the m-th coefficient of the $\boldsymbol{\beta}_n$ vector and $\boldsymbol{x}_{mk}$ is the m-th explanatory variable of the k-th measurement. Once the beta coefficients have been estimated from the data, we will be able to estimate the probability that any subsequent set of explanatory variables will result in any of the possible outcome categories.

Interpretations

There are various equivalent specifications and interpretations of logistic regression, which fit into different types of more general models, and allow different generalizations.

As a generalized linear model

The particular model used by logistic regression, which distinguishes it from standard linear regression and from other types of regression analysis used for binary-valued outcomes, is the way the probability of a particular outcome is linked to the linear predictor function:

$$\operatorname{logit}(\mathbb{E}[Y_i \mid x_{1,i},\ldots,x_{m,i}]) = \operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i}$$

Written using the more compact notation described above, this is:

$$\operatorname{logit}(\mathbb{E}[Y_i \mid \mathbf{X}_i]) = \operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol\beta \cdot \mathbf{X}_i$$

This formulation expresses logistic regression as a type of generalized linear model, which predicts variables with various types of probability distributions by fitting a linear predictor function of the above form to some sort of arbitrary transformation of the expected value of the variable.

The intuition for transforming using the logit function (the natural log of the odds) was explained above. It also has the practical effect of converting the probability (which is bounded to be between 0 and 1) to a variable that ranges over $(-\infty,+\infty)$ — thereby matching the potential range of the linear prediction function on the right side of the equation.

Both the probabilities p_i and the regression coefficients are unobserved, and the means of determining them is not part of the model itself. They are typically determined by some sort of optimization procedure, e.g. maximum likelihood estimation, that finds values that best fit the observed data (i.e. that give the most accurate predictions for the data already observed), usually subject to regularization conditions that seek to exclude unlikely values, e.g. extremely large values for any of the regression coefficients. The use of a regularization condition is equivalent to doing maximum a posteriori (MAP) estimation, an extension of maximum likelihood. (Regularization is most commonly done using a squared regularizing function, which is equivalent to placing a zero-mean Gaussian prior distribution on the coefficients, but other regularizers are also possible.) Whether or not regularization is used, it is usually not possible to find a closed-form solution; instead, an iterative numerical method must be used, such as iteratively reweighted least squares (IRLS) or, more commonly these days, a quasi-Newton method such as the L-BFGS method.[25]

The interpretation of the β_j parameter estimates is as the additive effect on the log of the odds for a unit change in the j the explanatory variable. In the case of a dichotomous explanatory variable, for instance, gender e^β is the estimate of the odds of having the outcome for, say, males compared with females.

An equivalent formula uses the inverse of the logit function, which is the logistic function, i.e.:

$$\mathbb{E}[Y_i \mid \mathbf{X}_i] = p_i = \operatorname{logit}^{-1}(\boldsymbol\beta \cdot \mathbf{X}_i) = \frac{1}{1+e^{-\boldsymbol\beta \cdot \mathbf{X}_i}}$$

The formula can also be written as a probability distribution (specifically, using a probability mass function):

$$\Pr(Y_i = y \mid \mathbf{X}_i) = {p_i}^y(1-p_i)^{1-y} = \left(\frac{e^{\boldsymbol\beta \cdot \mathbf{X}_i}}{1+e^{\boldsymbol\beta \cdot \mathbf{X}_i}}\right)^y \left(1-\frac{e^{\boldsymbol\beta \cdot \mathbf{X}_i}}{1+e^{\boldsymbol\beta \cdot \mathbf{X}_i}}\right)^{1-y} = \frac{e^{\boldsymbol\beta \cdot \mathbf{X}_i \cdot y}}{1+e^{\boldsymbol\beta \cdot \mathbf{X}_i}}$$

As a latent-variable model

The logistic model has an equivalent formulation as a latent-variable model. This formulation is common in the theory of discrete choice models and makes it easier to extend to certain more complicated models with multiple, correlated choices, as well as to compare logistic regression to the closely related probit model.

Imagine that, for each trial i, there is a continuous latent variable Y_i^* (i.e. an unobserved random variable) that is distributed as follows:

$$Y_i^\ast = \boldsymbol\beta \cdot \mathbf{X}_i + \varepsilon_i$$

where

$$\varepsilon_i \sim \operatorname{Logistic}(0,1)$$

i.e. the latent variable can be written directly in terms of the linear predictor function and an additive random error variable that is distributed according to a standard logistic distribution.

Then Y_i can be viewed as an indicator for whether this latent variable is positive:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^\ast > 0 \ \text{ i.e. } -\varepsilon_i < \boldsymbol\beta \cdot \mathbf{X}_i, \\ 0 & \text{otherwise.} \end{cases}$$

The choice of modeling the error variable specifically with a standard logistic distribution, rather than a general logistic distribution with the location and scale set to arbitrary values, seems restrictive, but in fact, it is not. It must be kept in mind that we can choose the regression coefficients ourselves, and very often can use them to offset changes in the parameters of the error variable's distribution. For example, a logistic error-variable distribution with a non-zero location parameter μ (which sets the mean) is equivalent to a distribution with a zero location parameter, where μ has been added to the intercept coefficient. Both situations produce the same value for Y_i^* regardless of settings of explanatory variables. Similarly, an arbitrary scale parameter s is equivalent to setting the scale parameter to 1 and then dividing all regression coefficients by s. In the latter case, the resulting value of Y_i^* will be smaller by a factor of s than in the former case, for all sets of explanatory variables — but critically, it will always remain on the same side of 0, and hence lead to the same Y_i choice.

(This predicts that the irrelevancy of the scale parameter may not carry over into more complex models where more than two choices are available.)

It turns out that this formulation is exactly equivalent to the preceding one, phrased in terms of the generalized linear model and without any latent variables. This can be shown as follows, using the fact that the cumulative distribution function (CDF) of the standard logistic distribution is the logistic function, which is the inverse of the logit function, i.e.

$$\Pr(\varepsilon_i < x) = \operatorname{logit}^{-1}(x)$$

Then:

$$\begin{aligned}
\Pr(Y_i = 1 \mid \mathbf{X}_i) &= \Pr(Y_i^\ast > 0 \mid \mathbf{X}_i) \\
&= \Pr(\boldsymbol\beta \cdot \mathbf{X}_i + \varepsilon_i > 0) \\
&= \Pr(\varepsilon_i > -\boldsymbol\beta \cdot \mathbf{X}_i) \\
&= \Pr(\varepsilon_i < \boldsymbol\beta \cdot \mathbf{X}_i) && \text{(because the logistic distribution is symmetric)} \\
&= \operatorname{logit}^{-1}(\boldsymbol\beta \cdot \mathbf{X}_i) \\
&= p_i && \text{(see above)}
\end{aligned}$$

This formulation—which is standard in discrete choice models—makes clear the relationship between logistic regression (the "logit model") and the probit model, which uses an error variable distributed according to a standard normal distribution instead of a standard logistic distribution. Both the logistic and normal distributions are symmetric with a basic unimodal, "bell curve" shape. The only difference is that the logistic distribution has somewhat heavier tails, which means that it is less sensitive to outlying data (and hence somewhat more robust to model mis-specifications or erroneous data).

Two-way latent-variable model

Yet another formulation uses two separate latent variables:

$$\begin{aligned}
Y_i^{0\ast} &= \boldsymbol\beta_0 \cdot \mathbf{X}_i + \varepsilon_0 \\
Y_i^{1\ast} &= \boldsymbol\beta_1 \cdot \mathbf{X}_i + \varepsilon_1
\end{aligned}$$

where

$$\begin{aligned}
\varepsilon_0 &\sim \operatorname{EV}_1(0,1) \\
\varepsilon_1 &\sim \operatorname{EV}_1(0,1)
\end{aligned}$$

where $EV_1(0,1)$ is a standard type-1 extreme value distribution: i.e.

$$\Pr(\varepsilon_0 = x) = \Pr(\varepsilon_1 = x) = e^{-x} e^{-e^{-x}}$$

Then

$$Y_i = \begin{cases} 1 & \text{if } Y_i^{1\ast} > Y_i^{0\ast}, \\ 0 & \text{otherwise.} \end{cases}$$

This model has a separate latent variable and a separate set of regression coefficients for each possible outcome of the dependent variable. The reason for this separation is that it makes it easy to extend logistic regression to multi-outcome categorical variables, as in the multinomial logit model. In such a model, it is natural to model each possible outcome using a different set of regression coefficients. It is also possible to motivate each of the separate latent variables as the theoretical utility associated with making the associated choice, and thus motivate logistic regression in terms of utility theory. (In terms of utility theory, a rational actor always chooses the choice with the greatest associated utility.) This is the approach taken by economists when formulating discrete choice models, because it both provides a theoretically strong foundation and facilitates intuitions about the model, which in turn makes it easy to consider various sorts of extensions. (See the example below.)

The choice of the type-1 extreme value distribution seems fairly arbitrary, but it makes the mathematics work out, and it may be possible to justify its use through rational choice theory.

It turns out that this model is equivalent to the previous model, although this seems non-obvious, since there are now two sets of regression coefficients and error variables, and the error variables have a different distribution. In fact, this model reduces directly to the previous one with the following substitutions:

$$\boldsymbol\beta = \boldsymbol\beta_1 - \boldsymbol\beta_0$$
$$\varepsilon = \varepsilon_1 - \varepsilon_0$$

An intuition for this comes from the fact that, since we choose based on the maximum of two values, only their difference matters, not the exact values — and this effectively removes one degree of freedom. Another critical fact is that the difference of two type-1 extreme-value-distributed variables is a logistic distribution, i.e. $\varepsilon = \varepsilon_1 - \varepsilon_0 \sim \operatorname{Logistic}(0,1).$ We can demonstrate the equivalent as follows:

$$\begin{aligned}
\Pr(Y_i = 1 \mid \mathbf{X}_i) &= \Pr\left(Y_i^{1\ast} > Y_i^{0\ast} \mid \mathbf{X}_i\right) \\
&= \Pr\left(Y_i^{1\ast} - Y_i^{0\ast} > 0 \mid \mathbf{X}_i\right) \\
&= \Pr\left(\boldsymbol\beta_1 \cdot \mathbf{X}_i + \varepsilon_1 - (\boldsymbol\beta_0 \cdot \mathbf{X}_i + \varepsilon_0) > 0\right) \\
&= \Pr\left((\boldsymbol\beta_1 \cdot \mathbf{X}_i - \boldsymbol\beta_0 \cdot \mathbf{X}_i) + (\varepsilon_1 - \varepsilon_0) > 0\right) \\
&= \Pr((\boldsymbol\beta_1 - \boldsymbol\beta_0) \cdot \mathbf{X}_i + (\varepsilon_1 - \varepsilon_0) > 0) \\
&= \Pr((\boldsymbol\beta_1 - \boldsymbol\beta_0) \cdot \mathbf{X}_i + \varepsilon > 0) & & \text{(substitute } \varepsilon \text{ as above)} \\
&= \Pr(\boldsymbol\beta \cdot \mathbf{X}_i + \varepsilon > 0) & & \text{(substitute } \boldsymbol\beta \text{ as above)} \\
&= \Pr(\varepsilon > -\boldsymbol\beta \cdot \mathbf{X}_i) & & \text{(now, same as above model)} \\
&= \Pr(\varepsilon < \boldsymbol\beta \cdot \mathbf{X}_i) \\
&= \operatorname{logit}^{-1}(\boldsymbol\beta \cdot \mathbf{X}_i) \\
&= p_i
\end{aligned}$$

Example

As an example, consider a province-level election where the choice is between a right-of-center party, a left-of-center party, and a secessionist party (e.g. the Parti Québécois, which wants Quebec to secede from Canada). We would then use three latent variables, one for each choice. Then, in accordance with utility theory, we can then interpret the latent variables as expressing the utility that results from making each of the choices. We can also interpret the regression coefficients as indicating the strength that the associated factor (i.e. explanatory variable) has in contributing to the utility — or more correctly, the amount by which a unit change in an explanatory variable changes the utility of a given choice. A voter might expect that the right-of-center party would lower taxes, especially on rich people. This would give low-income people no benefit, i.e. no change in utility (since they usually don't pay taxes); would cause moderate benefit (i.e. somewhat more money, or moderate utility increase) for middle-incoming people; would cause significant benefits for high-income people. On the other hand, the left-of-center party might be expected to raise taxes and offset it with increased welfare and other assistance for the lower and middle classes. This would cause significant positive benefit to low-income people, perhaps a weak benefit to middle-income people, and significant negative benefit to high-income people. Finally, the secessionist party would take no direct actions on the economy, but simply secede. A low-income or middle-income voter might expect basically no clear utility gain or loss from this, but a high-income voter might expect negative utility since he/she is likely to own companies, which will have a harder time doing business in such an environment and probably lose money.

These intuitions can be expressed as follows:

Estimated strength of regression coefficient for different outcomes (party choices) and different values of explanatory variables

	Center-right	Center-left	Secessionist
High-income	strong +	strong –	strong –
Middle-income	moderate +	weak +	none
Low-income	none	strong +	none

This clearly shows that

1. Separate sets of regression coefficients need to exist for each choice. When phrased in terms of utility, this can be seen very easily. Different choices have different effects on net utility; furthermore, the effects vary in complex ways that depend on the characteristics of each individual, so there need to be separate sets of coefficients for each characteristic, not simply a single extra per-choice characteristic.
2. Even though income is a continuous variable, its effect on utility is too complex for it to be treated as a single variable. Either it needs to be directly split up into ranges, or higher powers of income need to be added so that polynomial regression on income is effectively done.

As a "log-linear" model

Yet another formulation combines the two-way latent variable formulation above with the original formulation higher up without latent variables, and in the process provides a link to one of the standard formulations of the multinomial logit.

Here, instead of writing the logit of the probabilities p_i as a linear predictor, we separate the linear predictor into two, one for each of the two outcomes:

$$\ln \Pr(Y_i = 0) = \boldsymbol{\beta}_0 \cdot \mathbf{X}_i - \ln Z$$
$$\ln \Pr(Y_i = 1) = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i - \ln Z$$

Two separate sets of regression coefficients have been introduced, just as in the two-way latent variable model, and the two equations appear a form that writes the logarithm of the associated probability as a linear predictor, with an extra term $-\ln Z$ at the end. This term, as it turns out, serves as the normalizing factor ensuring that the result is a distribution. This can be seen by exponentiating both sides:

$$\Pr(Y_i = 0) = \frac{1}{Z} e^{\boldsymbol{\beta}_0 \cdot \mathbf{X}_i}$$

$$\Pr(Y_i = 1) = \frac{1}{Z} e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}$$

In this form it is clear that the purpose of Z is to ensure that the resulting distribution over Y_i is in fact a probability distribution, i.e. it sums to 1. This means that Z is simply the sum of all un-normalized probabilities, and by dividing each probability by Z, the probabilities become "normalized". That is:

$$Z = e^{\boldsymbol{\beta}_0 \cdot \mathbf{X}_i} + e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}$$

and the resulting equations are

$$\Pr(Y_i = 0) = \frac{e^{\boldsymbol{\beta}_0 \cdot \mathbf{X}_i}}{e^{\boldsymbol{\beta}_0 \cdot \mathbf{X}_i} + e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{e^{\boldsymbol{\beta}_0 \cdot \mathbf{X}_i} + e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}.$$

Or generally:

$$\Pr(Y_i = c) = \frac{e^{\boldsymbol\beta_c \cdot \mathbf{X}_i}}{\sum_h e^{\boldsymbol\beta_h \cdot \mathbf{X}_i}}$$

This shows clearly how to generalize this formulation to more than two outcomes, as in multinomial logit. This general formulation is exactly the softmax function as in

$$\Pr(Y_i = c) = \operatorname{softmax}(c, \boldsymbol\beta_0 \cdot \mathbf{X}_i, \boldsymbol\beta_1 \cdot \mathbf{X}_i, \dots).$$

To prove that this is equivalent to the previous model, we start by recognizing the above model is overspecified, in that $\Pr(Y_i = 0)$ and $\Pr(Y_i = 1)$ cannot be independently specified: rather $\Pr(Y_i = 0) + \Pr(Y_i = 1) = 1$ so knowing one automatically determines the other. As a result, the model is nonidentifiable, in that multiple combinations of $\boldsymbol\beta_0$ and $\boldsymbol\beta_1$ will produce the same probabilities for all possible explanatory variables. In fact, it can be seen that adding any constant vector to both of them will produce the same probabilities:

$$\begin{aligned}\Pr(Y_i = 1) &= \frac{e^{(\boldsymbol\beta_1 + \mathbf{C}) \cdot \mathbf{X}_i}}{e^{(\boldsymbol\beta_0 + \mathbf{C}) \cdot \mathbf{X}_i} + e^{(\boldsymbol\beta_1 + \mathbf{C}) \cdot \mathbf{X}_i}} \\ &= \frac{e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i} e^{\mathbf{C} \cdot \mathbf{X}_i}}{e^{\boldsymbol\beta_0 \cdot \mathbf{X}_i} e^{\mathbf{C} \cdot \mathbf{X}_i} + e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i} e^{\mathbf{C} \cdot \mathbf{X}_i}} \\ &= \frac{e^{\mathbf{C} \cdot \mathbf{X}_i} e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i}}{e^{\mathbf{C} \cdot \mathbf{X}_i} (e^{\boldsymbol\beta_0 \cdot \mathbf{X}_i} + e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i})} \\ &= \frac{e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i}}{e^{\boldsymbol\beta_0 \cdot \mathbf{X}_i} + e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i}}.\end{aligned}$$

As a result, we can simplify matters, and restore identifiability, by picking an arbitrary value for one of the two vectors. We choose to set $\boldsymbol\beta_0 = \mathbf{0}$. Then,

$$e^{\boldsymbol\beta_0 \cdot \mathbf{X}_i} = e^{\mathbf{0} \cdot \mathbf{X}_i} = 1$$

and so

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i}}{1 + e^{\boldsymbol\beta_1 \cdot \mathbf{X}_i}} = \frac{1}{1 + e^{-\boldsymbol\beta_1 \cdot \mathbf{X}_i}} = p_i$$

which shows that this formulation is indeed equivalent to the previous formulation. (As in the two-way latent variable formulation, any settings where $\boldsymbol\beta = \boldsymbol\beta_1 - \boldsymbol\beta_0$ will produce equivalent results.)

Most treatments of the multinomial logit model start out either by extending the "log-linear" formulation presented here or the two-way latent variable formulation presented above, since both clearly show the way that the model could be extended to multi-way outcomes. In general, the presentation with latent variables is more common in econometrics and political science, where discrete choice models and utility theory reign, while the "log-linear" formulation here is more common in computer science, e.g. machine learning and natural language processing.

As a single-layer perceptron

The model has an equivalent formulation

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}.$$

This functional form is commonly called a single-layer perceptron or single-layer artificial neural network. A single-layer neural network computes a continuous output instead of a step function. The derivative of p_i with respect to $X = (x_1, ..., x_k)$ is computed from the general form:

$$y = \frac{1}{1 + e^{-f(X)}}$$

where $f(X)$ is an analytic function in X. With this choice, the single-layer neural network is identical to the logistic regression model. This function has a continuous derivative, which allows it to be used in backpropagation. This function is also preferred because its derivative is easily calculated:

$$\frac{\mathrm{d}y}{\mathrm{d}X} = y(1-y)\frac{\mathrm{d}f}{\mathrm{d}X}.$$

In terms of binomial data

A closely related model assumes that each i is associated not with a single Bernoulli trial but with n_i independent identically distributed trials, where the observation Y_i is the number of successes observed (the sum of the individual Bernoulli-distributed random variables), and hence follows a binomial distribution:

$$Y_i \ \sim \operatorname{Bin}(n_i, p_i),\ \text{for } i = 1, \dots, n$$

An example of this distribution is the fraction of seeds (p_i) that germinate after n_i are planted.

In terms of expected values, this model is expressed as follows:

$$p_i = \mathbb{E}\left[\left.\frac{Y_i}{n_i}\,\right|\,\mathbf{X}_i\right],$$

so that

$$\operatorname{logit}\left(\mathbb{E}\left[\left.\frac{Y_i}{n_i}\,\right|\,\mathbf{X}_i\right]\right) = \operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol\beta \cdot \mathbf{X}_i\,,$$

Or equivalently:

$$\Pr(Y_i = y \mid \mathbf{X}_i) = \binom{n_i}{y} p_i^y (1-p_i)^{n_i-y} = \binom{n_i}{y}\left(\frac{1}{1+e^{-\boldsymbol\beta\cdot\mathbf{X}_i}}\right)^y \left(1 - \frac{1}{1+e^{-\boldsymbol\beta\cdot\mathbf{X}_i}}\right)^{n_i-y}.$$

This model can be fit using the same sorts of methods as the above more basic model.

Model fitting

Maximum likelihood estimation (MLE)

The regression coefficients are usually estimated using maximum likelihood estimation.[26][27] Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximize the likelihood function so an iterative process must be used instead; for example Newton's method. This process begins with a tentative solution, revises it slightly to see if it can be improved, and repeats this revision until no more improvement is made, at which point the process is said to have converged.[26]

In some instances, the model may not reach convergence. Non-convergence of a model indicates that the coefficients are not meaningful because the iterative process was unable to find appropriate solutions. A failure to converge may occur for a number of reasons: having a large ratio of predictors to cases, multicollinearity, sparseness, or complete separation.

- Having a large ratio of variables to cases results in an overly conservative Wald statistic (discussed below) and can lead to non-convergence. Regularized logistic regression is specifically intended to be used in this situation.
- Multicollinearity refers to unacceptably high correlations between predictors. As multicollinearity increases, coefficients remain unbiased but standard errors increase and the likelihood of model convergence decreases.[26] To detect multicollinearity amongst the predictors, one can conduct a linear regression analysis with the predictors of interest for the sole purpose of examining the tolerance statistic [26] used to assess whether multicollinearity is unacceptably high.
- Sparseness in the data refers to having a large proportion of empty cells (cells with zero counts). Zero cell counts are particularly problematic with categorical predictors. With continuous predictors, the model can infer values for the zero cell counts, but this is not the case with categorical predictors. The model will not converge with zero cell counts for categorical predictors because the natural logarithm of zero is an undefined value so that the final solution to the model cannot be

reached. To remedy this problem, researchers may collapse categories in a theoretically meaningful way or add a constant to all cells.[26]

- Another numerical problem that may lead to a lack of convergence is complete separation, which refers to the instance in which the predictors perfectly predict the criterion – all cases are accurately classified and the likelihood maximized with infinite coefficients. In such instances, one should re-examine the data, as there may be some kind of error.[2]
- One can also take semi-parametric or non-parametric approaches, e.g., via local-likelihood or nonparametric quasi-likelihood methods, which avoid assumptions of a parametric form for the index function and is robust to the choice of the link function (e.g., probit or logit).[28]

Iteratively reweighted least squares (IRLS)

Binary logistic regression ($y = 0$ or $y = 1$) can, for example, be calculated using *iteratively reweighted least squares* (IRLS), which is equivalent to maximizing the log-likelihood of a Bernoulli distributed process using Newton's method. If the problem is written in vector matrix form, with parameters $\mathbf{w}^T = [\beta_0, \beta_1, \beta_2, \ldots]$, explanatory variables $\mathbf{x}(i) = [1, x_1(i), x_2(i), \ldots]^T$ and expected value of the Bernoulli distribution $\mu(i) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}(i)}}$, the parameters $\mathbf{w}$ can be found using the following iterative algorithm:

$$\mathbf{w}_{k+1} = \left(\mathbf{X}^T \mathbf{S}_k \mathbf{X}\right)^{-1} \mathbf{X}^T \left(\mathbf{S}_k \mathbf{X} \mathbf{w}_k + \mathbf{y} - \boldsymbol{\mu}_k\right)$$

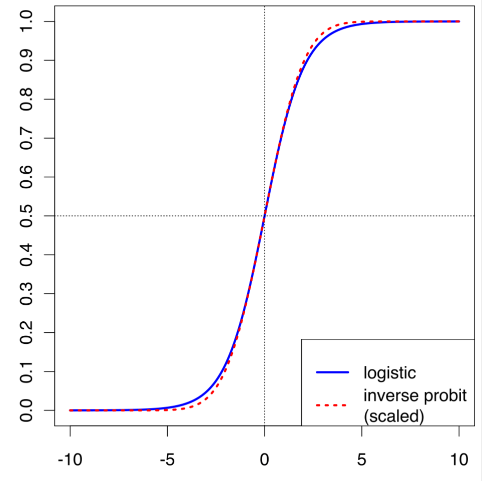
where $\mathbf{S} = \operatorname{\mathbf{diag}}(\mu(i)(1 - \mu(i)))$ is a diagonal weighting matrix, $\boldsymbol{\mu} = [\mu(1), \mu(2), \ldots]$ the vector of expected values,

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \ldots \\ 1 & x_1(2) & x_2(2) & \ldots \\ \vdots & \vdots & \vdots \end{bmatrix}$$

The regressor matrix and $\mathbf{y}(i) = [y(1), y(2), \ldots]^T$ the vector of response variables. More details can be found in the literature.[29]

Bayesian

In a Bayesian statistics context, prior distributions are normally placed on the regression coefficients, for example in the form of Gaussian distributions. There is no conjugate prior of the likelihood function in logistic regression. When Bayesian inference was performed analytically, this made the posterior distribution difficult to calculate except in very low dimensions. Now, though, automatic software such as OpenBUGS, JAGS, PyMC, Stan or Turing.jl allows these posteriors to be computed using simulation, so lack of conjugacy is not a concern. However, when the sample size or the number of parameters is large, full Bayesian simulation can be slow, and people often use approximate methods such as variational Bayesian methods and expectation propagation.



Comparison of logistic function with a scaled inverse probit function (i.e. the CDF of the normal distribution), comparing $\sigma(x)$ vs. $\Phi(\sqrt{\frac{\pi}{8}}x)$, which makes the slopes the same at the origin. This shows the heavier tails of the logistic distribution.

"Rule of ten"

Widely used, the "one in ten rule", states that logistic regression models give stable values for the explanatory variables if based on a minimum of about 10 events per explanatory variable (EPV); where *event* denotes the cases belonging to the less frequent category in the dependent variable. Thus a study designed to use k explanatory variables for an event (e.g. myocardial infarction) expected to occur in a proportion p of participants in the study will require a total of $10k/p$ participants. However, there is considerable debate about the reliability of this rule, which is based on simulation studies and lacks a secure theoretical underpinning.[30] According to some authors[31] the rule is overly conservative in some circumstances, with the authors stating, "If we (somewhat subjectively) regard confidence interval coverage less than 93 percent, type I error greater than 7 percent, or relative bias greater than 15 percent as problematic, our results indicate that problems are fairly frequent with 2–4 EPV, uncommon with 5–9 EPV, and still observed with 10–16 EPV. The worst instances of each problem were not severe with 5–9 EPV and usually comparable to those with 10–16 EPV".[32]

Others have found results that are not consistent with the above, using different criteria. A useful criterion is whether the fitted model will be expected to achieve the same predictive discrimination in a new sample as it appeared to achieve in the model development sample. For that criterion, 20 events per candidate variable may be required.[33] Also, one can argue that 96 observations are needed only to estimate the model's intercept precisely enough that the margin of error in predicted probabilities is ±0.1 with a 0.95 confidence level.[13]

Error and significance of fit

Deviance and likelihood ratio test — a simple case

In any fitting procedure, the addition of another fitting parameter to a model (e.g. the beta parameters in a logistic regression model) will almost always improve the ability of the model to predict the measured outcomes. This will be true even if the additional term has no predictive value, since the model will simply be "overfitting" to the noise in the data. The question arises as to whether the improvement gained by the addition of another fitting parameter is significant enough to recommend the inclusion of the additional term, or whether the improvement is simply that which may be expected from overfitting.

In short, for logistic regression, a statistic known as the deviance is defined which is a measure of the error between the logistic model fit and the outcome data. In the limit of a large number of data points, the deviance is chi-squared distributed, which allows a chi-squared test to be implemented in order to determine the significance of the explanatory variables.

Linear regression and logistic regression have many similarities. For example, in simple linear regression, a set of K data points (x_k, y_k) are fitted to a proposed model function of the form $y=b_0+b_1x$. The fit is obtained by choosing the b parameters which minimize the sum of the squares of the residuals (the squared error term) for each data point:

$$\varepsilon^2=\sum_{k=1}^{K}(b_0+b_1x_k-y_k)^2.$$

The minimum value which constitutes the fit will be denoted by $\hat{\varepsilon}^2$

The idea of a null model may be introduced, in which it is assumed that the x variable is of no use in predicting the y_k outcomes: The data points are fitted to a null model function of the form $y = b_0$ with a squared error term:

$$\varepsilon^2=\sum_{k=1}^{K}(b_0-y_k)^2.$$

The fitting process consists of choosing a value of b_0 which minimizes ε^2 of the fit to the null model, denoted by ε_φ^2 where the φ subscript denotes the null model. It is seen that the null model is optimized by $b_0=\overline{y}$ where $\overline{y}$ is the mean of the y_k values, and the optimized ε_φ^2 is:

$$\hat{\varepsilon}_\varphi^2=\sum_{k=1}^{K}(\overline{y}-y_k)^2$$

which is proportional to the square of the (uncorrected) sample standard deviation of the y_k data points.

We can imagine a case where the y_k data points are randomly assigned to the various x_k, and then fitted using the proposed model. Specifically, we can consider the fits of the proposed model to every permutation of the y_k outcomes. It can be shown that the optimized error of any of these fits will never be less than the optimum error of the null model, and that the difference between these minimum error will follow a chi-squared distribution, with degrees of freedom equal those of the proposed model minus those of the null model which, in this case, will be $2-1=1$. Using the chi-squared test, we may then estimate how many of these permuted sets of y_k will yield a minimum error less than or equal to the minimum error using the original y_k, and so we can estimate how significant an improvement is given by the inclusion of the x variable in the proposed model.

For logistic regression, the measure of goodness-of-fit is the likelihood function L, or its logarithm, the log-likelihood ℓ. The likelihood function L is analogous to the ε^2 in the linear regression case, except that the likelihood is maximized rather than minimized. Denote the maximized log-likelihood of the proposed model by $\hat{\ell}$.

In the case of simple binary logistic regression, the set of K data points are fitted in a probabilistic sense to a function of the form:

$$p(x) = \frac{1}{1+e^{-t}}$$

where $p(x)$ is the probability that $y = 1$. The log-odds are given by:

$$t = \beta_0 + \beta_1 x$$

and the log-likelihood is:

$$\ell = \sum_{k=1}^{K} \left(y_k \ln(p(x_k)) + (1-y_k)\ln(1-p(x_k))\right)$$

For the null model, the probability that $y = 1$ is given by:

$$p_\varphi(x) = \frac{1}{1+e^{-t_\varphi}}$$

The log-odds for the null model are given by:

$$t_\varphi = \beta_0$$

and the log-likelihood is:

$$\ell_\varphi = \sum_{k=1}^{K} \left(y_k \ln(p_\varphi) + (1-y_k)\ln(1-p_\varphi)\right)$$

Since we have $p_\varphi = \overline{y}$ at the maximum of L, the maximum log-likelihood for the null model is

$$\hat{\ell}_\varphi = K(\,\overline{y}\ln(\overline{y}) + (1-\overline{y})\ln(1-\overline{y}))$$

The optimum β_0 is:

$$\beta_0 = \ln\left(\frac{\overline{y}}{1-\overline{y}}\right)$$

where $\overline{y}$ is again the mean of the y_k values. Again, we can conceptually consider the fit of the proposed model to every permutation of the y_k and it can be shown that the maximum log-likelihood of these permutation fits will never be smaller than that of the null model:

$$\hat{\ell} \geq \hat{\ell}_\varphi$$

Also, as an analog to the error of the linear regression case, we may define the deviance of a logistic regression fit as:

$$D = \ln\left(\frac{\hat{L}^2}{\hat{L}_\varphi^2}\right) = 2(\hat{\ell} - \hat{\ell}_\varphi)$$

which will always be positive or zero. The reason for this choice is that not only is the deviance a good measure of the goodness of fit, it is also approximately chi-squared distributed, with the approximation improving as the number of data points (K) increases, becoming exactly chi-square distributed in the limit of an infinite number of data points. As in the case of linear regression, we may use this fact to estimate the probability that a random set of data points will give a better fit than the fit obtained by the proposed model, and so have an estimate how significantly the model is improved by including the x_k data points in the proposed model.

For the simple model of student test scores described above, the maximum value of the log-likelihood of the null model is $\hat{\ell}_\varphi = -13.8629\ldots$ The maximum value of the log-likelihood for the simple model is $\hat{\ell} = -8.02988\ldots$ so that the deviance is $D = 2(\hat{\ell} - \hat{\ell}_\varphi) = 11.6661\ldots$

Using the chi-squared test of significance, the integral of the chi-squared distribution with one degree of freedom from 11.6661... to infinity is equal to 0.00063649...

This effectively means that about 6 out of a 10,000 fits to random y_k can be expected to have a better fit (smaller deviance) than the given y_k and so we can conclude that the inclusion of the x variable and data in the proposed model is a very significant improvement over the null model. In other words, we reject the null hypothesis with $1 - D \approx 99.94\%$ confidence.

Goodness of fit summary

Goodness of fit in linear regression models is generally measured using R^2. Since this has no direct analog in logistic regression, various methods[34]:ch.21 including the following can be used instead.

Deviance and likelihood ratio tests

In linear regression analysis, one is concerned with partitioning variance via the sum of squares calculations – variance in the criterion is essentially divided into variance accounted for by the predictors and residual variance. In logistic regression analysis, deviance is used in lieu of a sum of squares calculations.[35] Deviance is analogous to the sum of squares calculations in linear regression[2] and is a measure of the lack of fit to the data in a logistic regression model.[35] When a "saturated" model is available (a model with a theoretically perfect fit), deviance is calculated by comparing a given model with the saturated model.[2] This computation gives the likelihood-ratio test:[2]

$$D = -2\ln\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}.$$

In the above equation, D represents the deviance and ln represents the natural logarithm. The log of this likelihood ratio (the ratio of the fitted model to the saturated model) will produce a negative value, hence the need for a negative sign. D can be shown to follow an approximate chi-squared distribution.[2] Smaller values indicate better fit as the fitted model deviates less from the saturated model. When assessed upon a chi-square distribution, nonsignificant chi-square values indicate very little unexplained variance and thus, good model fit. Conversely, a significant chi-square value indicates that a significant amount of the variance is unexplained.

When the saturated model is not available (a common case), deviance is calculated simply as −2·(log likelihood of the fitted model), and the reference to the saturated model's log likelihood can be removed from all that follows without harm.

Two measures of deviance are particularly important in logistic regression: null deviance and model deviance. The null deviance represents the difference between a model with only the intercept (which means "no predictors") and the saturated model. The model deviance represents the difference between a model with at least one predictor and the saturated model.[35] In this respect, the null model provides a baseline upon which to compare predictor models. Given that deviance is a measure of the difference between a given model and the saturated model, smaller values indicate better fit. Thus, to assess the contribution of a predictor or set of predictors, one can subtract the model deviance from the null deviance and assess the difference on a χ^2_{s-p}, chi-square distribution with degrees of freedom[2] equal to the difference in the number of parameters estimated.

Let

$$\begin{aligned} D_{\text{null}} &= -2\ln\frac{\text{likelihood of null model}}{\text{likelihood of the saturated model}} \\ D_{\text{fitted}} &= -2\ln\frac{\text{likelihood of fitted model}}{\text{likelihood of the saturated model}}. \end{aligned}$$

Then the difference of both is:

$$\begin{aligned}D_{\text{null}}-D_{\text{fitted}}&=-2\left(\ln\frac{\text{likelihood of null model}}{\text{likelihood of the saturated model}}-\ln\frac{\text{likelihood of fitted model}}{\text{likelihood of the saturated model}}\right)\\&=-2\ln\frac{\left(\dfrac{\text{likelihood of null model}}{\text{likelihood of the saturated model}}\right)}{\left(\dfrac{\text{likelihood of fitted model}}{\text{likelihood of the saturated model}}\right)}\\&=-2\ln\frac{\text{likelihood of the null model}}{\text{likelihood of fitted model}}.\end{aligned}$$

If the model deviance is significantly smaller than the null deviance then one can conclude that the predictor or set of predictors significantly improve the model's fit. This is analogous to the F-test used in linear regression analysis to assess the significance of prediction.[35]

Pseudo-R-squared

In linear regression the squared multiple correlation, R^2 is used to assess goodness of fit as it represents the proportion of variance in the criterion that is explained by the predictors.[35] In logistic regression analysis, there is no agreed upon analogous measure, but there are several competing measures each with limitations.[35][36]

Four of the most commonly used indices and one less commonly used one are examined on this page:

- Likelihood ratio $R^2{}_{\text{L}}$
- Cox and Snell $R^2{}_{\text{CS}}$
- Nagelkerke $R^2{}_{\text{N}}$
- McFadden $R^2{}_{\text{McF}}$
- Tjur $R^2{}_{\text{T}}$

Hosmer–Lemeshow test

The Hosmer–Lemeshow test uses a test statistic that asymptotically follows a χ^2 distribution to assess whether or not the observed event rates match expected event rates in subgroups of the model population. This test is considered to be obsolete by some statisticians because of its dependence on arbitrary binning of predicted probabilities and relative low power.[37]

Coefficient significance

After fitting the model, it is likely that researchers will want to examine the contribution of individual predictors. To do so, they will want to examine the regression coefficients. In linear regression, the regression coefficients represent the change in the criterion for each unit change in the predictor.[35] In logistic regression, however, the regression coefficients represent the change in the logit for each unit change in the predictor. Given that the logit is not intuitive, researchers are likely to focus on a predictor's effect on the exponential function of the regression coefficient – the odds ratio (see definition). In linear regression, the significance of a regression coefficient is assessed by computing a t test. In logistic regression, there are several different tests designed to assess the significance of an individual predictor, most notably the likelihood ratio test and the Wald statistic.

Likelihood ratio test

The likelihood-ratio test discussed above to assess model fit is also the recommended procedure to assess the contribution of individual "predictors" to a given model.[2][26][35] In the case of a single predictor model, one simply compares the deviance of the predictor model with that of the null model on a chi-square distribution with a single degree of freedom. If the predictor model has significantly smaller deviance (c.f. chi-square using the difference in degrees of freedom of the two models), then one can conclude that there is a significant association between the "predictor" and the outcome. Although some common statistical packages (e.g. SPSS) do provide likelihood ratio test statistics, without this computationally intensive test it would be more difficult to assess the contribution of individual predictors in the multiple logistic regression case. To assess the contribution of individual predictors one can enter the predictors hierarchically, comparing each new model with the previous to determine the contribution of each predictor.[35] There is some debate among statisticians about the appropriateness of so-called "stepwise" procedures. The fear is that they may not preserve nominal statistical properties and may become misleading.[38]

Wald statistic

Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the *t*-test in linear regression, is used to assess the significance of coefficients. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.[26]

$$W_j = \frac{\beta_j^2}{SE_{\beta_j}^2}$$

Although several statistical packages (e.g., SPSS, SAS) report the Wald statistic to assess the contribution of individual predictors, the Wald statistic has limitations. When the regression coefficient is large, the standard error of the regression coefficient also tends to be larger increasing the probability of Type-II error. The Wald statistic also tends to be biased when data are sparse.[35]

Case-control sampling

Suppose cases are rare. Then we might wish to sample them more frequently than their prevalence in the population. For example, suppose there is a disease that affects 1 person in 10,000 and to collect our data we need to do a complete physical. It may be too expensive to do thousands of physicals of healthy people in order to obtain data for only a few diseased individuals. Thus, we may evaluate more diseased individuals, perhaps all of the rare outcomes. This is also retrospective sampling, or equivalently it is called unbalanced data. As a rule of thumb, sampling controls at a rate of five times the number of cases will produce sufficient control data.[39]

Logistic regression is unique in that it may be estimated on unbalanced data, rather than randomly sampled data, and still yield correct coefficient estimates of the effects of each independent variable on the outcome. That is to say, if we form a logistic model from such data, if the model is correct in the general population, the β_j parameters are all correct except for β_0. We can correct β_0 if we know the true prevalence as follows:[39]

$$\widehat{\beta}_0^* = \widehat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

where π is the true prevalence and $\tilde{\pi}$ is the prevalence in the sample.

Discussion

Like other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Unlike ordinary linear regression, however, logistic regression is used for predicting dependent variables that take membership in one of a limited number of categories (treating the dependent variable in the binomial case as the outcome of a Bernoulli trial) rather than a continuous outcome. Given this difference, the assumptions of linear regression are violated. In particular, the residuals cannot be normally distributed. In addition, linear regression may make nonsensical predictions for a binary dependent variable. What is needed is a way to convert a binary variable into a continuous one that can take on any real value (negative or positive). To do that, binomial logistic regression first calculates the odds of the event happening for different levels of each independent variable, and then takes its logarithm to create a continuous criterion as a transformed version of the dependent variable. The logarithm of the odds is the logit of the probability, the logit is defined as follows:

$$\operatorname{logit} p = \ln \frac{p}{1-p} \quad \text{for } 0 < p < 1\,.$$

Although the dependent variable in logistic regression is Bernoulli, the logit is on an unrestricted scale.[2] The logit function is the link function in this kind of generalized linear model, i.e.

$$\operatorname{logit} \mathcal{E}(Y) = \beta_0 + \beta_1 x$$

Y is the Bernoulli-distributed response variable and x is the predictor variable; the β values are the linear parameters.

The logit of the probability of success is then fitted to the predictors. The predicted value of the logit is converted back into predicted odds, via the inverse of the natural logarithm – the exponential function. Thus, although the observed dependent variable in binary logistic regression is a 0-or-1 variable, the logistic regression estimates the odds, as a continuous variable, that the dependent variable is a 'success'. In some applications, the odds are all that is needed. In others, a specific yes-or-no prediction is needed for whether the dependent variable is or is not a 'success'; this categorical prediction can be based on the computed odds of success, with predicted odds above some chosen cutoff value being translated into a prediction of success.

Machine learning and cross-entropy loss function

In machine learning applications where logistic regression is used for binary classification, the MLE minimises the cross-entropy loss function.

Logistic regression is an important machine learning algorithm. The goal is to model the probability of a random variable Y being 0 or 1 given experimental data.[40]

Consider a generalized linear model function parameterized by θ,

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}} = \Pr(Y=1 \mid X;\theta)$$

Therefore,

$$\Pr(Y=0 \mid X;\theta) = 1 - h_\theta(X)$$

and since $Y \in \{0,1\}$, we see that $\Pr(y \mid X;\theta)$ is given by $\Pr(y \mid X;\theta) = h_\theta(X)^y(1-h_\theta(X))^{(1-y)}$. We now calculate the likelihood function assuming that all the observations in the sample are independently Bernoulli distributed,

$$\begin{aligned} L(\theta \mid y;x) &= \Pr(Y \mid X;\theta) \\ &= \prod_i \Pr(y_i \mid x_i;\theta) \\ &= \prod_i h_\theta(x_i)^{y_i}(1-h_\theta(x_i))^{(1-y_i)} \end{aligned}$$

Typically, the log likelihood is maximized,

$$N^{-1}\log L(\theta \mid y;x) = N^{-1}\sum_{i=1}^N \log \Pr(y_i \mid x_i;\theta)$$

which is maximized using optimization techniques such as gradient descent.

Assuming the (x,y) pairs are drawn uniformly from the underlying distribution, then in the limit of large *N*,

$$\begin{aligned} &\lim_{N\to+\infty} N^{-1}\sum_{i=1}^N \log\Pr(y_i \mid x_i;\theta) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(X=x,Y=y)\log\Pr(Y=y \mid X=x;\theta) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(X=x,Y=y)\left(-\log\frac{\Pr(Y=y \mid X=x)}{\Pr(Y=y \mid X=x;\theta)} + \log\Pr(Y=y \mid X=x)\right) \\ &= -D_{\text{KL}}(Y \parallel Y_\theta) - H(Y \mid X) \end{aligned}$$

where $H(Y \mid X)$ is the conditional entropy and D_{KL} is the Kullback–Leibler divergence. This leads to the intuition that by maximizing the log-likelihood of a model, you are minimizing the KL divergence of your model from the maximal entropy distribution. Intuitively searching for the model that makes the fewest assumptions in its parameters.

Comparison with linear regression

Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between the dependent and independent variables) from those of linear regression. In particular, the key differences between these two models can be seen in the following two features of logistic regression. First, the conditional distribution $y \mid x$ is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the **probability** of particular outcomes rather than the outcomes themselves.

Alternatives

A common alternative to the logistic model (logit model) is the probit model, as the related names suggest. From the perspective of generalized linear models, these differ in the choice of link function: the logistic model uses the logit function (inverse logistic function), while the probit model uses the probit function (inverse error function). Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.[41] Other sigmoid functions or error distributions can be used instead.

Logistic regression is an alternative to Fisher's 1936 method, linear discriminant analysis.[42] If the assumptions of linear discriminant analysis hold, the conditioning can be reversed to produce logistic regression. The converse is not true, however, because logistic regression does not require the multivariate normal assumption of discriminant analysis.[43]

The assumption of linear predictor effects can easily be relaxed using techniques such as spline functions.[13]

History

A detailed history of the logistic regression is given in Cramer (2002). The logistic function was developed as a model of population growth and named "logistic" by Pierre François Verhulst in the 1830s and 1840s, under the guidance of Adolphe Quetelet; see Logistic function § History for details.[44] In his earliest paper (1838), Verhulst did not specify how he fit the curves to the data.[45][46] In his more detailed paper (1845), Verhulst determined the three parameters of the model by making the curve pass through three observed points, which yielded poor predictions.[47][48]

The logistic function was independently developed in chemistry as a model of autocatalysis (Wilhelm Ostwald, 1883).[49] An autocatalytic reaction is one in which one of the products is itself a catalyst for the same reaction, while the supply of one of the reactants is fixed. This naturally gives rise to the logistic equation for the same reason as population growth: the reaction is self-reinforcing but constrained.

The logistic function was independently rediscovered as a model of population growth in 1920 by Raymond Pearl and Lowell Reed, published as Pearl & Reed (1920), which led to its use in modern statistics. They were initially unaware of Verhulst's work and presumably learned about it from L. Gustave du Pasquier, but they gave him little credit and did not adopt his terminology.[50] Verhulst's priority was acknowledged and the term "logistic" revived by Udny Yule in 1925 and has been followed since.[51] Pearl and Reed first applied the model to the population of the United States, and also initially fitted the curve by making it pass through three points; as with Verhulst, this again yielded poor results.[52]

In the 1930s, the probit model was developed and systematized by Chester Ittner Bliss, who coined the term "probit" in Bliss (1934), and by John Gaddum in Gaddum (1933), and the model fit by maximum likelihood estimation by Ronald A. Fisher in Fisher (1935), as an addendum to Bliss's work. The probit model was principally used in bioassay, and had been preceded by earlier work dating to 1860; see Probit model § History. The probit model influenced the subsequent development of the logit model and these models competed with each other.[53]

The logistic model was likely first used as an alternative to the probit model in bioassay by Edwin Bidwell Wilson and his student Jane Worcester in Wilson & Worcester (1943).[54] However, the development of the logistic model as a general alternative to the probit model was principally due to the work of Joseph Berkson over many decades, beginning in Berkson (1944), where he coined "logit", by analogy with "probit", and continuing through Berkson (1951) and following years.[55] The logit model was initially dismissed as inferior to the probit model, but "gradually achieved an equal footing with the probit",[56] particularly between 1960 and 1970. By 1970, the logit model achieved parity with the probit model in use in statistics journals and thereafter surpassed it. This relative popularity was due to the adoption of the logit outside of bioassay, rather than displacing the probit within bioassay, and its informal use in practice; the logit's popularity is credited to the logit model's computational simplicity, mathematical properties, and generality, allowing its use in varied fields.[3]

Various refinements occurred during that time, notably by David Cox, as in Cox (1958).[4]

The multinomial logit model was introduced independently in Cox (1966) and Theil (1969), which greatly increased the scope of application and the popularity of the logit model.[57] In 1973 Daniel McFadden linked the multinomial logit to the theory of discrete choice, specifically Luce's choice axiom, showing that the multinomial logit followed from the assumption of independence of irrelevant alternatives and interpreting odds of alternatives as relative preferences;[58] this gave a theoretical foundation for the logistic regression.[57]

Extensions

There are large numbers of extensions:

- Multinomial logistic regression (or **multinomial logit**) handles the case of a multi-way categorical dependent variable (with unordered values, also called "classification"). The general case of having dependent variables with more than two values is termed *polytomous regression*.
- Ordered logistic regression (or **ordered logit**) handles ordinal dependent variables (ordered values).
- Mixed logit is an extension of multinomial logit that allows for correlations among the choices of the dependent variable.
- An extension of the logistic model to sets of interdependent variables is the conditional random field.
- Conditional logistic regression handles matched or stratified data when the strata are small. It is mostly used in the analysis of observational studies.

See also

$\sqrt{x}$ ***Mathematics portal***

- Logistic function
- Discrete choice
- Jarrow–Turnbull model
- Limited dependent variable
- Multinomial logit model
- Ordered logit
- Hosmer–Lemeshow test
- Brier score
- mlpack - contains a C++ implementation of logistic regression
- Local case-control sampling
- Logistic model tree

References

1. Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". *JAMA*. **316** (5): 533–4. doi:10.1001/jama.2016.7653 (https://doi.org/10.1001%2Fjama.2016.7653). ISSN 0098-7484 (https://search.worldcat.org/issn/0098-7484). OCLC 6823603312 (https://search.worldcat.org/oclc/6823603312). PMID 27483067 (https://pubmed.ncbi.nlm.nih.gov/27483067).
2. Hosmer, David W.; Lemeshow, Stanley (2000). *Applied Logistic Regression* (2nd ed.). Wiley. ISBN 978-0-471-35632-5.
3. Cramer 2002, p. 10–11.
4. Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". *Biometrika*. **54** (1/2): 167–178. doi:10.2307/2333860 (https://doi.org/10.2307%2F2333860). JSTOR 2333860 (https://www.jstor.org/stable/2333860).
5. Cramer 2002, p. 8.
6. Boyd, C. R.; Tolson, M. A.; Copes, W. S. (1987). "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score" (https://doi.org/10.1097%2F00005373-198704000-00005). *The Journal of Trauma*. **27** (4): 370–378. doi:10.1097/00005373-198704000-00005 (https://doi.org/10.1097%2F00005373-198704000-00005). PMID 3106646 (https://pubmed.ncbi.nlm.nih.gov/3106646).

7. Kologlu, M.; Elker, D.; Altun, H.; Sayek, I. (2001). "Validation of MPI and PIA II in two different groups of patients with secondary peritonitis". *Hepato-Gastroenterology*. **48** (37): 147–51. PMID 11268952 (https://pubmed.ncbi.nlm.nih.gov/11268952).
8. Biondo, S.; Ramos, E.; Deiros, M.; Ragué, J. M.; De Oca, J.; Moreno, P.; Farran, L.; Jaurrieta, E. (2000). "Prognostic factors for mortality in left colonic peritonitis: A new scoring system". *Journal of the American College of Surgeons*. **191** (6): 635–42. doi:10.1016/S1072-7515(00)00758-4 (https://doi.org/10.1016%2FS1072-7515%2800%2900758-4). PMID 11129812 (https://pubmed.ncbi.nlm.nih.gov/11129812).
9. Marshall, J. C.; Cook, D. J.; Christou, N. V.; Bernard, G. R.; Sprung, C. L.; Sibbald, W. J. (1995). "Multiple organ dysfunction score: A reliable descriptor of a complex clinical outcome". *Critical Care Medicine*. **23** (10): 1638–52. doi:10.1097/00003246-199510000-00007 (https://doi.org/10.1097%2F00003246-199510000-00007). PMID 7587228 (https://pubmed.ncbi.nlm.nih.gov/7587228).
10. Le Gall, J. R.; Lemeshow, S.; Saulnier, F. (1993). "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study". *JAMA*. **270** (24): 2957–63. doi:10.1001/jama.1993.03510240069035 (https://doi.org/10.1001%2Fjama.1993.03510240069035). PMID 8254858 (https://pubmed.ncbi.nlm.nih.gov/8254858).
11. David A. Freedman (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. p. 128.
12. Truett, J; Cornfield, J; Kannel, W (1967). "A multivariate analysis of the risk of coronary heart disease in Framingham". *Journal of Chronic Diseases*. **20** (7): 511–24. doi:10.1016/0021-9681(67)90082-3 (https://doi.org/10.1016%2F0021-9681%2867%2990082-3). PMID 6028270 (https://pubmed.ncbi.nlm.nih.gov/6028270).
13. Harrell, Frank E. (2015). *Regression Modeling Strategies*. Springer Series in Statistics (2nd ed.). New York; Springer. doi:10.1007/978-3-319-19425-7 (https://doi.org/10.1007%2F978-3-319-19425-7). ISBN 978-3-319-19424-0.
14. M. Strano; B.M. Colosimo (2006). "Logistic regression analysis for experimental determination of forming limit diagrams". *International Journal of Machine Tools and Manufacture*. **46** (6): 673–682. doi:10.1016/j.ijmachtools.2005.07.005 (https://doi.org/10.1016%2Fj.ijmachtools.2005.07.005).
15. Palei, S. K.; Das, S. K. (2009). "Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach". *Safety Science*. **47**: 88–96. doi:10.1016/j.ssci.2008.01.002 (https://doi.org/10.1016%2Fj.ssci.2008.01.002).
16. Berry, Michael J.A (1997). *Data Mining Techniques For Marketing, Sales and Customer Support*. Wiley. p. 10.
17. Mesa-Arango, Rodrigo; Hasan, Samiul; Ukkusuri, Satish V.; Murray-Tuite, Pamela (February 2013). "Household-Level Model for Hurricane Evacuation Destination Type Choice Using Hurricane Ivan Data" (https://ascelibrary.org/doi/10.1061/%28ASCE%29NH.1527-6996.0000083). *Natural Hazards Review*. **14** (1): 11–20. Bibcode:2013NHRev..14...11M (https://ui.adsabs.harvard.edu/abs/2013NHRev..14...11M). doi:10.1061/(ASCE)NH.1527-6996.0000083 (https://doi.org/10.1061%2F%28ASCE%29NH.1527-6996.0000083). ISSN 1527-6988 (https://search.worldcat.org/issn/1527-6988).
18. Wibbenmeyer, Matthew J.; Hand, Michael S.; Calkin, David E.; Venn, Tyron J.; Thompson, Matthew P. (June 2013). "Risk Preferences in Strategic Wildfire Decision Making: A Choice Experiment with U.S. Wildfire Managers" (https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2012.01894.x). *Risk Analysis*. **33** (6): 1021–1037. Bibcode:2013RiskA..33.1021W (https://ui.adsabs.harvard.edu/abs/2013RiskA..33.1021W). doi:10.1111/j.1539-6924.2012.01894.x (https://doi.org/10.1111%2Fj.1539-6924.2012.01894.x). ISSN 0272-4332 (https://search.worldcat.org/issn/0272-4332). PMID 23078036 (https://pubmed.ncbi.nlm.nih.gov/23078036). S2CID 45282555 (https://api.semanticscholar.org/CorpusID:45282555).
19. Lovreglio, Ruggiero; Borri, Dino; dell'Olio, Luigi; Ibeas, Angel (2014-02-01). "A discrete choice model based on random utilities for exit choice in emergency evacuations" (https://www.sciencedirect.com/science/article/pii/S0925753513002294). *Safety Science*. **62**: 418–426. doi:10.1016/j.ssci.2013.10.004 (https://doi.org/10.1016%2Fj.ssci.2013.10.004). ISSN 0925-7535 (https://search.worldcat.org/issn/0925-7535).
20. "Logistic Regression" (https://www.mastersindatascience.org/learning/machine-learning-algorithms/logistic-regression/). *CORP-MIDS1 (MDS)*. Retrieved 2024-03-16.
21. Neyman, J.; Pearson, E. S. (1933), "On the problem of the most efficient tests of statistical hypotheses" (http://www.stats.org.uk/statistical-inference/NeymanPearson1933.pdf) (PDF), *Philosophical Transactions of the Royal Society of London A*, **231** (694–706): 289–337, Bibcode:1933RSPTA.231..289N (https://ui.adsabs.harvard.edu/abs/1933RSPTA.231..289N), doi:10.1098/rsta.1933.0009 (https://doi.org/10.1098%2Frsta.1933.0009), JSTOR 91247 (https://www.jstor.org/stable/91247)
22. "How to Interpret Odds Ratio in Logistic Regression?" (https://stats.idre.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/). Institute for Digital Research and Education.
23. Everitt, Brian (1998). *The Cambridge Dictionary of Statistics* (https://archive.org/details/cambridgediction00ever_0). Cambridge, UK New York: Cambridge University Press. ISBN 978-0-521-59346-5.
24. For example, the indicator function in this case could be defined as $\Delta(n,y)=1-(y-n)^2$
25. Malouf, Robert (2002). "A comparison of algorithms for maximum entropy parameter estimation" (https://dl.acm.org/citation.cfm?id=1118871). *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*. pp. 49–55. doi:10.3115/1118853.1118871 (https://doi.org/10.3115%2F1118853.1118871).
26. Menard, Scott W. (2002). *Applied Logistic Regression* (2nd ed.). SAGE. ISBN 978-0-7619-2208-7.
27. Gourieroux, Christian; Monfort, Alain (1981). "Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models". *Journal of Econometrics*. **17** (1): 83–97. doi:10.1016/0304-4076(81)90060-9 (https://doi.org/10.1016%2F0304-4076%2881%2990060-9).

28. Park, Byeong U.; Simar, Léopold; Zelenyuk, Valentin (2017). "Nonparametric estimation of dynamic discrete choice models for time series data" (https://espace.library.uq.edu.au/view/UQ:415620/UQ415620_OA.pdf) (PDF). *Computational Statistics & Data Analysis*. **108**: 97–120. doi:10.1016/j.csda.2016.10.024 (https://doi.org/10.1016%2Fj.csda.2016.10.024).
29. Murphy, Kevin P. (2012). *Machine Learning – A Probabilistic Perspective*. The MIT Press. p. 245. ISBN 978-0-262-01802-9.
30. Van Smeden, M.; De Groot, J. A.; Moons, K. G.; Collins, G. S.; Altman, D. G.; Eijkemans, M. J.; Reitsma, J. B. (2016). "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5122171). *BMC Medical Research Methodology*. **16** (1): 163. doi:10.1186/s12874-016-0267-3 (https://doi.org/10.1186%2Fs12874-016-0267-3). PMC 5122171 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5122171). PMID 27881078 (https://pubmed.ncbi.nlm.nih.gov/27881078).
31. Peduzzi, P; Concato, J; Kemper, E; Holford, TR; Feinstein, AR (December 1996). "A simulation study of the number of events per variable in logistic regression analysis" (https://doi.org/10.1016%2Fs0895-4356%2896%2900236-3). *Journal of Clinical Epidemiology*. **49** (12): 1373–9. doi:10.1016/s0895-4356(96)00236-3 (https://doi.org/10.1016%2Fs0895-4356%2896%2900236-3). PMID 8970487 (https://pubmed.ncbi.nlm.nih.gov/8970487).
32. Vittinghoff, E.; McCulloch, C. E. (12 January 2007). "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression" (https://doi.org/10.1093%2Faje%2Fkwk052). *American Journal of Epidemiology*. **165** (6): 710–718. doi:10.1093/aje/kwk052 (https://doi.org/10.1093%2Faje%2Fkwk052). PMID 17182981 (https://pubmed.ncbi.nlm.nih.gov/17182981).
33. van der Ploeg, Tjeerd; Austin, Peter C.; Steyerberg, Ewout W. (2014). "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4289553). *BMC Medical Research Methodology*. **14**: 137. doi:10.1186/1471-2288-14-137 (https://doi.org/10.1186%2F1471-2288-14-137). PMC 4289553 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4289553). PMID 25532820 (https://pubmed.ncbi.nlm.nih.gov/25532820).
34. Greene, William N. (2003). *Econometric Analysis* (Fifth ed.). Prentice-Hall. ISBN 978-0-13-066189-0.
35. Cohen, Jacob; Cohen, Patricia; West, Steven G.; Aiken, Leona S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Routledge. ISBN 978-0-8058-2223-6.
36. Allison, Paul D. "Measures of fit for logistic regression" (https://support.sas.com/resources/papers/proceedings14/1485-2014.pdf) (PDF). Statistical Horizons LLC and the University of Pennsylvania.
37. Hosmer, D.W. (1997). "A comparison of goodness-of-fit tests for the logistic regression model". *Stat Med*. **16** (9): 965–980. doi:10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.3.co;2-f (https://doi.org/10.1002%2F%28sici%291097-0258%2819970515%2916%3A9%3C965%3A%3Aaid-sim509%3E3.3.co%3B2-f). PMID 9160492 (https://pubmed.ncbi.nlm.nih.gov/9160492).
38. Harrell, Frank E. (2010). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer. ISBN 978-1-4419-2918-1.
39. https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/classification.pdf slide 16
40. Ng, Andrew (2000). "CS229 Lecture Notes" (http://akademik.bahcesehir.edu.tr/~tevfik/courses/cmp5101/cs229-notes1.pdf) (PDF). *CS229 Lecture Notes*: 16–19.
41. Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models* (http://data.princeton.edu/wws509/notes/). pp. Chapter 3, page 45.
42. Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). *An Introduction to Statistical Learning* (http://www-bcf.usc.edu/~gareth/ISL/). Springer. p. 6.
43. Pohar, Maja; Blas, Mateja; Turk, Sandra (2004). "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study" (https://www.researchgate.net/publication/229021894). *Metodološki Zvezki*. **1** (1).
44. Cramer 2002, pp. 3–5.
45. Verhulst, Pierre-François (1838). "Notice sur la loi que la population poursuit dans son accroissement" (https://books.google.com/books?id=8GsEAAAAYAAJ) (PDF). *Correspondance Mathématique et Physique*. **10**: 113–121. Retrieved 3 December 2014.
46. Cramer 2002, p. 4, "He did not say how he fitted the curves."
47. Verhulst, Pierre-François (1845). "Recherches mathématiques sur la loi d'accroissement de la population" (http://gdz.sub.uni-goettingen.de/dms/load/img/?PPN=PPN129323640_0018&DMDID=dmdlog7) [Mathematical Researches into the Law of Population Growth Increase]. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*. **18**. Retrieved 2013-02-18.
48. Cramer 2002, p. 4.
49. Cramer 2002, p. 7.
50. Cramer 2002, p. 6.
51. Cramer 2002, p. 6–7.
52. Cramer 2002, p. 5.
53. Cramer 2002, p. 7–9.
54. Cramer 2002, p. 9.
55. Cramer 2002, p. 8, "As far as I can see the introduction of the logistics as an alternative to the normal probability function is the work of a single person, Joseph Berkson (1899–1982), ..."
56. Cramer 2002, p. 11.

57. Cramer 2002, p. 13.
58. McFadden, Daniel (1973). "Conditional Logit Analysis of Qualitative Choice Behavior" (https://web.archive.org/web/20181127110612/https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf) (PDF). In P. Zarembka (ed.). *Frontiers in Econometrics*. New York: Academic Press. pp. 105–142. Archived from the original (https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf) (PDF) on 2018-11-27. Retrieved 2019-04-20.

Sources

- Berkson, Joseph (1944). "Application of the Logistic Function to Bio-Assay". *Journal of the American Statistical Association*. **39** (227): 357–365. doi:10.1080/01621459.1944.10500699 (https://doi.org/10.1080%2F01621459.1944.10500699). JSTOR 2280041 (https://www.jstor.org/stable/2280041).
- Berkson, Joseph (1951). "Why I Prefer Logits to Probits". *Biometrics*. **7** (4): 327–339. doi:10.2307/3001655 (https://doi.org/10.2307%2F3001655). ISSN 0006-341X (https://search.worldcat.org/issn/0006-341X). JSTOR 3001655 (https://www.jstor.org/stable/3001655).
- Bliss, C. I. (1934). "The Method of Probits". *Science*. **79** (2037): 38–39. Bibcode:1934Sci....79...38B (https://ui.adsabs.harvard.edu/abs/1934Sci....79...38B). doi:10.1126/science.79.2037.38 (https://doi.org/10.1126%2Fscience.79.2037.38). PMID 17813446 (https://pubmed.ncbi.nlm.nih.gov/17813446). "These arbitrary probability units have been called 'probits'."
- Cox, David R. (1958). "The regression analysis of binary sequences (with discussion)". *J R Stat Soc B*. **20** (2): 215–242. doi:10.1111/j.2517-6161.1958.tb00292.x (https://doi.org/10.1111%2Fj.2517-6161.1958.tb00292.x). JSTOR 2983890 (https://www.jstor.org/stable/2983890).
- Cox, David R. (1966). "Some procedures connected with the logistic qualitative response curve". In F. N. David (ed.). *Research Papers in Probability and Statistics (Festschrift for J. Neyman)*. London: Wiley. pp. 55–71.
- Cramer, J. S. (2002). *The origins of logistic regression* (https://papers.tinbergen.nl/02119.pdf) (PDF) (Technical report). Vol. 119. Tinbergen Institute. pp. 167–178. doi:10.2139/ssrn.360300 (https://doi.org/10.2139%2Fssrn.360300).
  - Published in: Cramer, J. S. (2004). "The early origins of the logit model". *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*. **35** (4): 613–626. doi:10.1016/j.shpsc.2004.09.003 (https://doi.org/10.1016%2Fj.shpsc.2004.09.003).
- Fisher, R. A. (1935). "The Case of Zero Survivors in Probit Assays" (https://archive.today/20140430203018/http://ebooks.adelaide.edu.au/dspace/handle/2440/15223). *Annals of Applied Biology*. **22**: 164–165. doi:10.1111/j.1744-7348.1935.tb07713.x (https://doi.org/10.1111%2Fj.1744-7348.1935.tb07713.x). Archived from the original (https://ebooks.adelaide.edu.au/dspace/handle/2440/15223) on 2014-04-30.
- Gaddum, John H. (1933). *Reports on Biological Standards: Methods of biological assay depending on a quantal response. III*. H.M. Stationery Office. OCLC 808240121 (https://search.worldcat.org/oclc/808240121).
- Theil, Henri (1969). "A Multinomial Extension of the Linear Logit Model". *International Economic Review*. **10** (3): 251–59. doi:10.2307/2525642 (https://doi.org/10.2307%2F2525642). JSTOR 2525642 (https://www.jstor.org/stable/2525642).
- Pearl, Raymond; Reed, Lowell J. (June 1920). "On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1084522). *Proceedings of the National Academy of Sciences*. **6** (6): 275–288. Bibcode:1920PNAS....6..275P (https://ui.adsabs.harvard.edu/abs/1920PNAS....6..275P). doi:10.1073/pnas.6.6.275 (https://doi.org/10.1073%2Fpnas.6.6.275). PMC 1084522 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1084522). PMID 16576496 (https://pubmed.ncbi.nlm.nih.gov/16576496).
- Wilson, E.B.; Worcester, J. (1943). "The Determination of L.D.50 and Its Sampling Error in Bio-Assay" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1078563). *Proceedings of the National Academy of Sciences of the United States of America*. **29** (2): 79–85. Bibcode:1943PNAS...29...79W (https://ui.adsabs.harvard.edu/abs/1943PNAS...29...79W). doi:10.1073/pnas.29.2.79 (https://doi.org/10.1073%2Fpnas.29.2.79). PMC 1078563 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1078563). PMID 16588606 (https://pubmed.ncbi.nlm.nih.gov/16588606).
- Agresti, Alan. (2002). *Categorical Data Analysis*. New York: Wiley-Interscience. ISBN 978-0-471-36093-3.
- Amemiya, Takeshi (1985). "Qualitative Response Models" (https://books.google.com/books?id=0bzGQE14CwEC&pg=PA267). *Advanced Econometrics*. Oxford: Basil Blackwell. pp. 267–359. ISBN 978-0-631-13345-2.
- Balakrishnan, N. (1991). *Handbook of the Logistic Distribution*. Marcel Dekker, Inc. ISBN 978-0-8247-8587-1.
- Gouriéroux, Christian (2000). "The Simple Dichotomy" (https://books.google.com/books?id=dE2prs_U0QMC&pg=PA6). *Econometrics of Qualitative Dependent Variables*. New York: Cambridge University Press. pp. 6–37. ISBN 978-0-521-58985-7.
- Greene, William H. (2003). *Econometric Analysis, fifth edition*. Prentice Hall. ISBN 978-0-13-066189-0.
- Hilbe, Joseph M. (2009). *Logistic Regression Models*. Chapman & Hall/CRC Press. ISBN 978-1-4200-7575-5.
- Hosmer, David (2013). *Applied logistic regression*. Hoboken, New Jersey: Wiley. ISBN 978-0-470-58247-3.
- Howell, David C. (2010). *Statistical Methods for Psychology, 7th ed*. Belmont, CA; Thomson Wadsworth. ISBN 978-0-495-59786-5.
- Peduzzi, P.; J. Concato; E. Kemper; T.R. Holford; A.R. Feinstein (1996). "A simulation study of the number of events per variable in logistic regression analysis" (https://doi.org/10.1016%2Fs0895-4356%2896%2900236-3). *Journal of Clinical Epidemiology*. **49** (12): 1373–1379. doi:10.1016/s0895-4356(96)00236-3 (https://doi.org/10.1016%2Fs0895-4356%2896%2900236-3). PMID 8970487 (https://pubmed.ncbi.nlm.nih.gov/8970487).

- Berry, Michael J.A.; Linoff, Gordon (1997). *Data Mining Techniques For Marketing, Sales and Customer Support*. Wiley.

External links

- Media related to Logistic regression at Wikimedia Commons
- Econometrics Lecture (topic: Logit model) (https://www.youtube.com/watch?v=JvioZoK1f4o&t=64m48s) on YouTube by Mark Thoma
- Logistic Regression tutorial (http://www.omidrouhani.com/research/logisticregression/html/logisticregression.htm)
- mlelr (https://czep.net/stat/mlelr.html): software in C for teaching purposes

Retrieved from "https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=1325573741"