LOGISTIC REGRESSION

Joseph M. Hilbe

Arizona State University

Logistic regression is the most common method used to model binary response data. When the response is binary, it typically takes the form of 1/0, with 1 generally indicating a success and 0 a failure. However, the actual values that 1 and 0 can take vary widely, depending on the purpose of the study. For example, for a study of the odds of failure in a school setting, 1 may have the value of *fail*, and 0 of *not-fail*, or pass. The important point is that 1 indicates the foremost subject of interest for which a binary response study is designed. Modeling a binary response variable using normal linear regression introduces substantial bias into the parameter estimates. The standard linear model assumes that the response and error terms are normally or Gaussian distributed, that the variance, σ^2, is constant across observations, and that observations in the model are independent. When a binary variable is modeled using this method, the first two of the above assumptions are violated. Analogical to the normal regression model being based on the Gaussian probability distribution function (*pdf*), a binary response model is derived from a Bernoulli distribution, which is a subset of the binomial *pdf* with the binomial denominator taking the value of 1. The Bernoulli *pdf* may be expressed as:

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{1}$$

Binary logistic regression derives from the canonical form of the Bernoulli distribution. The Bernoulli *pdf* is a member of the exponential family of probability distributions, which has properties allowing for a much easier estimation of its parameters than traditional Newton-Raphson-based maximum likelihood estimation (MLE) methods.

In 1972 Nelder and Wedderbrun discovered that it was possible to construct a single algorithm for estimating models based on the exponential family of distributions. The algorithm was termed **Generalized linear models** (GLM), and became a standard method to estimate binary response models such as logistic, probit, and complimentary-loglog regression, count response models such as Poisson and negative binomial regression, and continuous response models such as gamma and inverse Gaussian regression. The standard normal model, or Gaussian regression, is also a generalized linear model, and may be estimated under its algorithm. The form of the exponential distribution appropriate for generalized linear models may be expressed as

$$f(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/\alpha(\phi) + c(y_i;\phi)\}, \tag{2}$$

with θ representing the link function, $\alpha(\phi)$ the scale parameter, $b(\theta)$ the cumulant, and $c(y;\phi)$ the normalization term, which guarantees that the probability function sums to 1. The link, a monotonically increasing function, linearizes the relationship of the expected mean and explanatory predictors. The scale, for binary and count models, is constrained to a value of 1, and the cumulant is used to calculate the model mean and variance functions. The mean is given as the first derivative of the cumulant with respect to θ, $b'(\theta)$; the variance is given as the second derivative, $b''(\theta)$. Taken together, the above four terms define a specific *GLM* model.

We may structure the Bernoulli distribution (Eq 3) into exponential family form (Eq 2) as:

$$f(y_i; \pi_i) = \exp\{y_i \ln(\pi_i/(1-\pi_i)) + \ln(1-\pi_i)\}. \tag{3}$$

The link function is therefore $\ln(\pi/(1-\pi))$, and cumulant $-\ln(1-\pi)$ or $\ln(1/(1-\pi))$. For the Bernoulli, π is defined as the probability of a success. The first derivative of the cumulant is π, the second derivative, $\pi(1-\pi)$. These two values are, respectively, the mean and variance functions of the Bernoulli *pdf*. Recalling that the logistic model is the canonical form of the distribution, meaning that it is the form that is directly derived from the *pdf*, the values expressed in Eq 3, and the values we gave for the mean and variance, are the values for the logistic model.

Estimation of statistical models using the *GLM* algorithm, as well as *MLE*, are both based on the log-likelihood function. The likelihood is simply a re-parameterization of the *pdf* which seeks to estimate π, for example, rather than y. The log-likelihood is formed from the likelihood by taking the natural log of the function, allowing summation across observations during the estimation process rather than multiplication.

The traditional *GLM* symbol for the mean, μ, is typically substituted for π, when *GLM* is used to estimate a logistic model. In that form, the log-likelihood function for the binary-logistic model is given as:

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i/(1-\mu_i)) + \ln(1-\mu_i)\}, \tag{4}$$

or

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) + (1-y_i)\ln(1-\mu_i)\}. \tag{5}$$

The Bernoulli-logistic log-likelihood function is essential to logistic regression. When *GLM* is used to estimate logistic models, many software algorithms use the deviance rather than the log-likelihood function as the basis of convergence. The deviance, which can be used as a goodness-of-fit statistic, is defined as twice the difference of the saturated log-likelihood and model log-likelihood. For logistic model, the deviance is expressed as

$$D = 2\sum_{i=1}^{n} \{y_i \ln(y_i/\mu_i) + (1-y_i)\ln((1-y_i)/(1-\mu_i))\}. \tag{6}$$

Whether estimated using maximum likelihood techniques or as *GLM*, the value of μ for each observation in the model is calculated on the basis of the linear predictor,

$x'\beta$. For the normal model, the predicted fit, $\hat{y}$, is identical to $x'\beta$, the right side of Equation 7. However, for logistic models, the response is expressed in terms of the link function, $\ln(\mu_i/(1-\mu_i))$. We have, therefore,

$$x_i'\beta = \ln(\mu_i/(1-\mu_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \tag{7}$$

The value of μ_i, for each observation in the logistic model, is calculated as

$$\mu_i = 1/(1+\exp(-x_i'\beta)) = \exp(x_i'\beta)/(1+\exp(x_i'\beta)). \tag{8}$$

The functions to the right of μ are commonly used ways of expressing the logistic inverse link function, which converts the linear predictor to the fitted value. For the logistic model, μ is a probability.

When logistic regression is estimated using a Newton-Raphson type of MLE algorithm, the log-likelihood function as parameterized to $x'\beta$ rather than μ. The estimated fit is then determined by taking the first derivative of the log-likelihood function with respect to β, setting it to zero, and solving. The first derivative of the log-likelihood function is commonly referred to as the gradient, or score function. The second derivative of the log-likelihood with respect to β produces the Hessian matrix, from which the standard errors of the predictor parameter estimates are derived. The logistic gradient and hessian functions are given as

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i) x_i \tag{9}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \left\{ x_i x_i' \mu_i (1-\mu_i) \right\} \tag{10}$$

One of the primary values of using the logistic regression model is the ability to interpret the exponentiated parameter estimates as odds ratios. Note that the link function is the log of the odds of μ, $\ln(\mu/(1-\mu))$, where the odds are understood as the success of μ over its failure, $1-\mu$. The log-odds is commonly referred to as the *logit* function. An example will help clarify the relationship, as well as the interpretation of the odds ratio.

We use data from the 1912 Titanic accident, comparing the odds of survival for adult passengers to children. A tabulation of the data is given as:

	Age (Child vs Adult)		
Survived	child	adults	Total
no	52	765	817
yes	57	442	499
Total	109	1,207	1,316

The odds of survival for adult passengers is 442/765, or 0.578. The odds of survival for children is 57/52, or 1.096. The ratio of the odds of survival for adults to the odds of survival for children is (442/765)/(57/52), or 0.52709552. This value is referred to as the *odds ratio*, or ratio of the two component odds relationships. Using a logistic regression procedure to estimate the odds ratio of age produces the following results

survived	Odds Ratio	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
age	.5270955	.1058718	-3.19	0.001	.3555642	.7813771

With $1 =$ *adult* and $0 =$ *child*, the estimated odds ratio may be interpreted as:

The odds of an adult surviving were about half the odds of a child surviving.

By inverting the estimated odds ratio above, we may conclude that children had $[1/.527 \sim 1.9]$ some 90% — or nearly two times – greater odds of surviving than did adults.

For continuous predictors, a one-unit increase in a predictor value indicates the change in odds expressed by the displayed odds ratio. For example, if age was recorded as a continuous predictor in the Titanic data, and the odds ratio was calculated as 1.015, we would interpret the relationship as:

The odds of surviving is one and a half percent greater for each increasing year of age.

Non-exponentiated logistic regression parameter estimates are interpreted as log-odds relationships, which carry little meaning in ordinary discourse. Logistic models are typically interpreted in terms of odds ratios, unless a researcher is interested in estimating predicted probabilities for given patterns of model covariates; i.e., in estimating μ.

Logistic regression may also be used for grouped or proportional data. For these models the response consists of a numerator, indicating the number of successes (1s) for a specific covariate pattern, and the denominator (m), the number of observations having the specific covariate pattern. The response y/m is binomially distributed as:

$$f(y_i; \pi_i, m_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}, \tag{11}$$

with a corresponding log-likelihood function expressed as

$$L(\mu_i; y_i, m_i) = \sum_{i=1}^{n} \left\{ y_i \ln(\mu_i/(1 - \mu_i)) + m_i \ln(1 - \mu_i) + \binom{m_i}{y_i} \right\}. \tag{12}$$

Taking derivatives of the cumulant, $-m_i \ln(1 - \mu_i)$, as we did for the binary response model, produces a mean of $\mu_i = m_i \pi_i$ and variance, $\mu_i(1 - \mu_i/m_i)$.

Consider the data below:

y	cases	x_1	x_2	x_3
1	3	1	0	1
1	1	1	1	1
2	2	0	0	1
0	1	0	1	1
2	2	1	0	0
0	1	0	1	0

y indicates the number of times a specific pattern of covariates is successful. *Cases* is the binomial denominator. The first observation in the table informs us that there are three cases having predictor values of $x_1 = 1, x_2 = 0$, and $x_3 = 1$. Of those three cases, one has a value of y equal to 1, the other two have values of 0. All current commercial software applications estimate this type of logistic model using GLM methodology.

y	Odds Ratio	OIM Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
x_1	1.186947	1.769584	0.11	0.908	.0638853	22.05271
x_2	.2024631	.3241584	-1.00	0.318	.0087803	4.668551
x_3	.5770337	.9126937	-0.35	0.728	.025993	12.8099

The data in the above table may be restructured so that it is in individual observation format, rather than grouped. The new table would have ten observations, having the same logic as described. Modeling would result in identical parameter estimates. It is not uncommon to find an individual-based data set of, for example, 10,000 observations, being grouped into 10-15 rows or observations as above described. Data in tables is nearly always expressed in grouped format.

Logistic models are subject to a variety of fit tests. Some of the more popular tests include the Hosmer-Lemeshow goodness-of-fit test, ROC analysis, various information criteria tests, link tests, and residual analysis. The Hosmer-Lemeshow test, once well used, is now only used with caution. The test is heavily influenced by the manner in which tied data is classified. Comparing observed with expected probabilities across levels, it is now preferred to construct tables of risk having different numbers of levels. If there is consistency in results across tables, then the statistic is more trustworthy.

Information criteria tests, e.g., Akaike information Criteria (AIC) and Bayesian Information Criteria (BIC) are the most used of this type of test. Information tests are comparative, with lower values indicating the preferred model. Recent research indicates that AIC and BIC both are biased when data is correlated to any degree. Statisticians have attempted to develop enhancements of these two tests, but have not been entirely successful. The best advice is to use several different types of tests, aiming for consistency of results.

Several types of residual analyses are typically recommended for logistic models. The references below provide extensive discussion of these methods, together with appropriate caveats. However, it appears well established that m-asymptotic residual analyses is most appropriate for logistic models having no continuous predictors. m-asymptotics is based on grouping observations with the same covariate pattern, in a similar manner to the grouped or binomial logistic regression discussed earlier. The Hilbe (2009) and Hosmer and Lemeshow (2000) references below provide guidance on how best to construct and interpret this type of residual.

Logistic models have been expanded to include categorical responses, e.g. proportional odds models and multinomial logistic regression. They have also been enhanced to include the modeling of panel and correlated data, e.g. generalized estimating equations, fixed and random effects, and mixed effects logistic models.

Finally, exact logistic regression models have recently been developed to allow the modeling of perfectly predicted data, as well as small and unbalanced datasets. In these cases, logistic models which are estimated using GLM or full maximum likelihood will not converge. Exact models employ entirely different methods of estimation, based on large numbers of permutations. Refer to *exact methods* or *computer intensive methods* for an elaboration of this technique.

References

[1] Collett, D. (2003). *Modeling Binary Regression.* 2nd edition, London: Chapman & Hall/CRC.

[2] Cox, D. R., E. J. Snell, (1989). *Analysis of Binary Data.* 2nd edition, London: Chapman & Hall.

[3] Hardin, J. W., J.M. Hilbe (2007). *Generalized Linear Models and Extensions.* 2nd edition, College Station: Stata Press.

[4] Hilbe, J. M. (2009) *Logistic Regression Models.* Boca Raton, FL: Chapman & Hall/CRC.

[5] Hosmer, D., S. Lemeshow (2000). *Applied Logistic Regression.* 2nd edition, New York: Wiley.

[6] Kleinbaum, D. G. (1994) *Logistic Regression; A Self-Teaching Guide.* New York: Springer.

[7] Long, J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage.

[8] McCullagh, P., J. Nelder (1989). *Generalized Linear Models.* 2nd edition, London: Chapman & Hall.