Generalized Linear Models

Link Function

The logistic equation is stated in terms of the probability that $Y = 1$, which is π, and the probability that $Y = 0$, which is 1 - π.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

The left-hand side of the equation represents the logit transformation, which takes the natural log of the ratio of the probability that Y is equal to 1 compared to the probability that it is not equal to one. As we know, the probability, π, is just the mean of the Y values, assuming 0,1 coding, which is often expressed as μ. The logit transformation could then be written in terms of the mean rather than the probability,

$$\ln\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta X$$

The transformation of the mean represents a link to the central tendency of the distribution, sometimes called the *location*, one of the important defining aspects of any given probability distribution. The log transformation represents a kind of *link function* (often *canonical link function*)[1] that is sometimes given more generally as $g(.)$, with the letter g used as an arbitrary name for a mathematical function and the use of the "." within the parentheses to suggest that any variable, value, or function (the *argument*) could be placed within. For logistic regression, this is known as the *logit link function*. The right hand side of the equation, $\alpha + \beta X$, is the familiar equation for the regression line and represents a linear combination of the parameters for the regression.

The concept of this logistic link function can generalized to any other distribution, with the simplest, most familiar case being the ordinary least squares or linear regression model. For the linear regression model, the link function is called the *identity link function*, because no transformation is needed to get from the linear regression parameters on the right-hand side of the equation to the normal distribution.

$$\hat{Y} = E(Y) = g\left(E(Y)\right) = g(\mu) = \alpha + \beta X$$

I give four equivalent terms here, $\hat{Y}$, $E(Y)$, $g\left(E(Y)\right)$, and $g(\mu)$, just to illustrate all of the different notations that might be used to express the linear regression model—we don't need them all. The expected value $E(Y)$ or mean, μ, of the response is plugged into $g(.)$ function because the predicted value is the expected value for Y when X is equal to some particular value.

The general concept that we can use a variety of link functions on the left-hand side of the equation and still keep the linear parameters on the right is referred to as the *generalized linear model* (Nelder & Wedderburn, 1972). The term should not be confused with the term "general linear model" used to refer generally to regression and ANOVA and their equivalence, which are special cases of the generalized linear model.[2]

---

[1] Technically, there is a distinction between a link function generally speaking and a canonical link function (see Agresti, 2015, pp. 3,123,142). A canonical link function is one in which transforms the mean, $\mu = E(y_i)$, to the natural exponential (location) parameter for the exponential family of distributions (e.g., normal, binomial, Poisson, gamma). The canonical link function is the most commonly used link form in generalized linear models.

[2] GLM is sometimes used for either generalized linear model or general linear model. GLIM is another abbreviation that is used only for the generalized linear model.

Some Generalized Linear Modeling Link Functions

Link type	Natural/Canonical Parameter Transformation	Example Application
Normal/Identity (OLS)	μ	
Log	$\ln \mu$	Poisson loglinear model for counts
Inverse	$1/\mu$	Regression with gamma distributed response
Square root	$\sqrt{\mu}$	Gamma distributed response increasing variance
Logit	$\ln\left(\pi/(1-\pi)\right)$	Binary and ordinal logistic regression
Probit	$\frac{1}{\sqrt{2\pi^*}}\int_{-\infty}^{\pi} e^{-z^2/2}dz$ (normal or Gaussian)	Binary and ordinal probit regressions
Log-log (also known as complementary log-log, Weibull)	$\ln\left[-\ln(1-\pi)\right]$	Survival analysis
Poisson	$\frac{\mu^y}{Y!}e^{-\mu}$	Regression of count response (equidispersion)
Negative binomial	$\frac{\Gamma(y_i+\omega)}{y!\Gamma(\omega)}\bullet\frac{\mu_i^{y_i}\omega^{\omega}}{(\mu_i+\omega)^{\mu_i+\omega}}$	Regression of count response

Random Component

The second component of the generalized linear model is the probability distribution associated with the with a particular type of variable—the distribution that the errors from the model are expected to follow. There are a number of distributions that fall under the exponential family of distributions, whose densities are all described with specific mathematical equations for the shape(s) of the distribution (see the overhead "Some Members of the Exponential Distribution Family"). All of the exponential family of distributions can be expressed in a very general form that has two parameters, the *natural* or *canonical parameter* (the location which is some function of the mean) and the *variance parameter*.[3] For ordinary least squares, it is the normal distribution. For logistic regression, it is the logistic distribution. Several other distributions are commonly used, including the Poisson for count variables, the inverse normal for the probit model, or the log-normal and log-logistic distributions used in survival analysis.

Generalized linear models are specified by indicating both the link function and the residual distribution. Sometimes a particular link is always used with a particular distribution, but sometimes there may be several possible distributions for a certain link. Maximum likelihood estimation is used for generalized linear models, with the usual significance test for overall model fit and coefficients—Wald, likelihood ratio, score tests (see Agresti, 2015, Chapter 4 for details on estimation and standard errors). Software packages, such as SPSS (*Genlin*), SAS (*PROC GENMOD*), and *glm* in R, allow users to specify link functions and distributions for a particular analytic circumstance.

Probit Regression Model

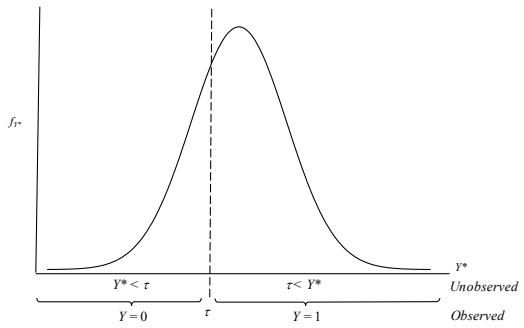
The logit link function is a fairly simple transformation of the prediction curve and also provides odds ratios, both features that make it popular among researchers. Another possibility when the dependent variable is dichotomous is *probit regression*. For some dichotomous variables, one can argue that the dependent variable is a proxy for a variable that is really continuous. Take for example our hypothetical child age and divorce study. Divorce might be the dichotomy that is ultimately observed, but there may be an underlying propensity toward divorce falling along some continuum related to marital satisfaction. Only when the propensity exceeds some threshold value on the continuum do we observe 1 (divorce) on

[3] The two parameters for a distribution are usually given using a function notation, such as $f(y; \mu,\sigma)$, where μ is the natural parameter and σ is the variance parameter.

the binary variable instead 0 (married). This underlying continuous variable is often called a *latent* or unobserved variable,[4] and the probit link function can be conceptualized as the link between the linear combination of parameters on the right-hand side to some unobserved continuum on the left-hand side of the generalized linear model. Below, I use Y^* (the Greek letter eta, η, is often used instead) to refer to the latent predicted score.

$$Y^* = \alpha + \beta X$$

The figure below illustrates the concept, using Y as the observed score, Y^*, and τ (tau) as the threshold.



Because the Y^* distribution is assumed to be normal, the unstandardized probit coefficients represent a change in the z-score for Y^* for each unit change in X. You can think about this as a partially standardized solution, with the dependent but not the independent variable standardized. The next step is the standardize X, to obtain a fully standardized solution, which provides a familiar metric and a convenient magnitude of effect for the association between each predictor and the response.

If the true underlying variable we are predicting is continuous, we can assume the errors are normally distributed. The probit regression model uses a (inverse) normal distribution link for a binary variable instead of the logit link, where $Y^* = \Phi^{-1}[\pi]$. The -1 superscript refers to the inverse of the cdf to correspond with the cumulative probability that Y is equal to 1.

$$\Phi^{-1}(\pi) = \frac{1}{\sqrt{2\pi^*}} \int_{-\infty}^{\pi} e^{-z^2/2} dz$$

π^* is used in the above for mathematical constant to distinguish it from the probability. Φ^{-1} is the *probit*, and, like the simpler logit, it connects the linear model with the expected probability.

Using the inverse normal function (in a statistical package or spreadsheet) for an observed probability returns a z-score on the normal distribution. The complementary function to the inverse normal cdf is the normal cdf, Φ, which can be used to transform a z-score back to a probability (i.e., the underlying mathematical transformation behind conversions obtained from a z statistical table).

$$\Phi^{-1}(\pi_i) = \alpha + \beta X_i \qquad\qquad \pi_i = \Phi(\alpha + \beta X_i)$$

The transformation, thus, represents the translation of Y to Y^* and back in the figure above. Similarly, values from the logistic model can be used to return an expected probability for a given value of X from

[4] Not really referring to the same concept as the term "latent variable" used in structural equation modeling, where a latent variable is estimated by a set of observed indicators assessing the same construct.

the model, except is simpler mathematical transformation to obtain the predicted probability from the cdf, $e^{\alpha+\beta X} / \left(1+e^{\alpha+\beta X}\right)$.

The probit regression is related to *polychoric* correlations, which does not require designation of an explanatory and response variable. Polychoric correlations were originally developed by Karl Pearson (1901) to correct for the loss of information in the usual Pearson correlations due to categorization of a continuous variable (see Olsson, 1979; MacCallum, Zhang, Preacher, & Rucker, 2002). The term polychoric is used more generally, but *Tetrachoric* correlations are a special case of polychoric correlations involving only binary variables, and *polyserial* correlations are those involving the correlation between a binary and a continuous variable (see also the "Analysis of Ordinal Contingency Tables" handout for more information). The concept of Y^* is the same as that invoked to conceptualize probit analysis, where the polychoric correlation represents the correlation between two Y^* variable. The variable Y^* is a true value that is not observed but leads to the observed response of Y, which is binary or ordinal.

Probit Regression vs. Logistic Regression

Probit regression and logistic regression can both model a binary dependent response. The difference between the two is just the link and error distributions assumed. As we know from the binomial test, with reasonably large n the normal and binomial distributions are very similar. Here is a picture of the cdf for the normal, standard logistic (usual, raw logistic), and the standardized logistic (assuming a standard deviation equal to $\pi^* / \sqrt{3}$).

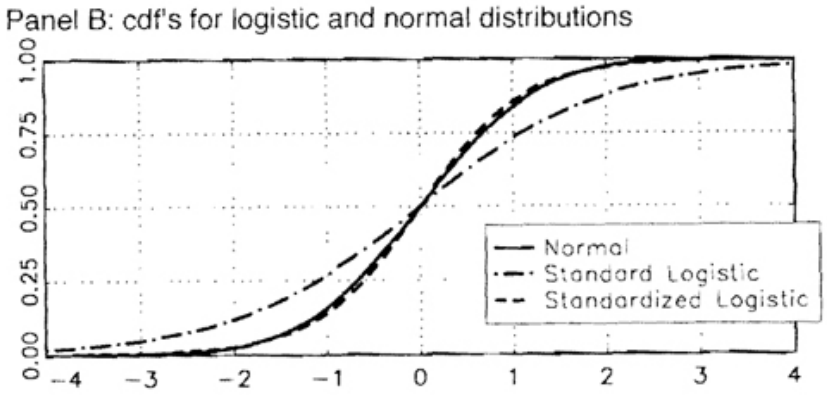


Figure 3.3. Normal and Logistic Distributions

From J. S. Long, 1997, p. 43

As this figure suggests, probit and logistic regression models nearly always produce the same statistical result. The unstandardized coefficient estimates from the two modeling approaches are on a different scale, given the different link functions (logit vs. probit), although the logistic coefficients tend to be approximately 1.7 larger than probit coefficients.[5] Different disciplines tend to use one more frequently than the other, although logistic regression is by far the most common. Logistic regression provides odds ratios, and probit models produce easily defined standardized coefficients.

[5] The difference tends to vary between about 1.6 and 1.8 and depends on the overall proportion of the outcome. This difference in units is connected to the variances of the logistic and normal probability distributions, where the proportion and the variance for binary variables are interdependent. The standardized logistic variance, which is $\left(\pi^*\right)^2 / 3 \approx 1.81$, leads to a cdf that is very close to the normal cdf, but this is based on the average across all values of X.

References and Further Reading

Agresti, A. (2015). *Foundations of linear and generalized linear models*. New York: John Wiley & Sons.
Dunteman, G. H., & Ho, M. H. R. (2005). *An introduction to generalized linear models (Vol. 145)*. Thousand Oaks, CA: Sage Publications.
Fox, J. (2008). *Applied regression analysis and generalized linear models, second edition*. Sage
Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical. Society Series A, 135*, 370-384.
Pearson, K. (1901). Mathematical contributions to the theory of evolution. X. Supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *197*, 443-459.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460