

Link Functions and the Generalized Linear Model

The Logit Link Function

Logistic regression can be thought of as consisting of a mathematical transformation of a standard regression model. Remember that one solution to outliers or heteroscedasticity problems is to transform X or Y or both by taking the square root or the log etc. The transformation used in logistic regression is a transformation of the predicted scores of Y (\hat{Y}), which is different. The transformation in logistic regression is called the *logit* transformation (so sometimes logistic is referred to as a *logit model*). Instead of using \hat{Y} , the log of the probabilities is used.

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X$$

The primary reason why the logit transformation function is used is that the residuals will not be normally distributed and they cannot be constant across values of X. Because Y has only two possible values 0 and 1, the residuals have only two possible values for each X. With only two possible values, the residuals cannot be normally distributed. Moreover, the best line to describe the relationship between X and Y is not likely to be linear, but rather an S-shape.

Instead of a normal distribution of errors, we assume the errors are logistically distributed. The basis of the logit link function is the cumulative frequency distribution, called a *cumulative distribution function* or CDF, that describes the distribution of the residuals. The *binomial* CDF is used because there are two possible outcomes.

The Probit Link Function

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 widget study. Whether a business succeeds or fails is really a matter or degree—some are more successful than others, some are more miserable failures than others. So, theoretically, continuous variables may underlie many dichotomous variables. That underlying continuous variable is often called a *latent* variable (related to the idea of polychoric correlations; Olsson, 1979). If we think about a regression analysis predicting the underlying latent variable, we have a probit analysis. Below, I use Y^* (the Greek letter eta, η , is sometimes used instead) to refer to the latent predicted score.

$$Y^* = B_0 + B_1x$$

If the true underlying variable we are predicting is continuous, we can assume the errors are normally distributed as we do in practice with OLS. In this case, instead of using the binomial CDF, we can use a link function based on the normal CDF. Remember, though, that because the relationship between X and Y is not linear, we cannot just use OLS. The following formula describes probit function (viewer discretion is advised!).

$$\Phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X} \exp\left(-\frac{1}{2}Z^2\right) dZ$$

I won't bother to define all of these symbols, since you don't need to memorize this. The capital phi, Φ , is used to designate the *probit* link function. Instead of the log transformation of the predicted scores, the probit transformation is used.

Generalized Linear Models

Using this same idea about link functions, we can transform any predicted curve to conform to different assumptions about the form of the relationship and the error distribution (Nelder & Wedderburn, 1972). We can think of all of these as part of the same *generalized linear model*. To denote the predicted curve for continuous variables, I use μ for the expected value of Y (usually referred to as $E(Y_i)$) at a particular value of X . For the predicted curve of dichotomous variables, I use π , for the expected probability, $E(P)$. The following formulas describe the link functions for different distributions:

Log link: $\ln \mu$

Inverse link: $\frac{1}{\mu}$

Square root link: $\sqrt{\mu}$

Logit link: $\ln\left(\frac{\pi}{1-\pi}\right)$

Probit link: $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X} \exp\left(-\frac{1}{2}Z^2\right) dZ$

Log-log link: $\ln[-\ln(1-\pi)]$

Poisson: $\frac{\mu^y}{Y!} e^{-\mu}$

Negative binomial: $\frac{\Gamma(y_i + \omega)}{y! \Gamma(\omega)} \cdot \frac{\mu_i^{y_i} \omega^\omega}{(\mu_i + \omega)^{\mu_i + \omega}}$

The log-log link function is for extreme asymmetric distributions and is sometimes used in complementary log-log regression model applications including survival analysis applications. The Poisson and negative binomial links are for regression models with count data (see Regression Models for Count Data handout). Generalized linear models are extremely useful because the regression model can be "linearized" to accommodate any form of predictive relationship and a variety of error distributions. Software packages, such as SPSS (*Genlin*) and SAS (*PROC GENMOD*), allow users to specify link functions and distributions for a particular analytic circumstance.

References and Suggested Reading

- Agresti, A. (2002). *Categorical Data Analysis, Second Edition*. New York: Wiley.
- Dunteman, G.H. and Moon-Ho, R.H. (2006). *An Introduction to Generalized Linear Models*. (Quantitative Applications in the Social Sciences). Thousand Oaks, CA: Sage
- Fox, J. (2008). *Applied regression analysis and generalized linear models, second edition*. Los Angeles, CA: Sage.
- Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- O'Connell, A.A. (2006). *Logistic Regression Models for Ordinal Response Variables*. (Quantitative Applications in the Social Sciences). Thousand Oaks, CA: Sage
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460.