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. Keep in mind that the transformation used in logistic regression is a transformation of the predicted scores ($\hat{Y}$), which is different from transforming the dependent variable (Y). The transformation in logistic regression is called the *logit* transformation (so sometimes logistic is referred to as a *logit model* if there is a binary independent variable). Instead of using $\hat{Y}$, the natural log of the probabilities is used.

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

The primary reason why the logit transformation function is used is that the best line to describe the relationship between X and Y is not likely to be linear, but rather an S-shape. Secondly, the conditional distribution of Y (i.e., the residuals) will differ from the conditional distribution when the outcome is continuous. 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 residuals determined in this way, they are unlikely to be normally distributed. Moreover, 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*.[1] 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 the binary variable instead 0 (married). This underlying continuous variable is often called a *latent response variable*.[2] 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.

$$Y^* = \Phi^{-1}[\hat{p}] = B_0 + B_1x$$

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.[3] In this case, instead of using the logistic cdf as with logistic regression, we can use a link function based on the normal cdf. The symbol $\Phi^{-1}[\hat{p}]^{-1}$ is used to designate the *probit* transformation of the predicted values—the link function. The -1 superscript refers to the inverse of the cdf to correspond with the cumulative probability that Y is equal to 1. The following formula describes the normal cdf (viewer discretion advised).

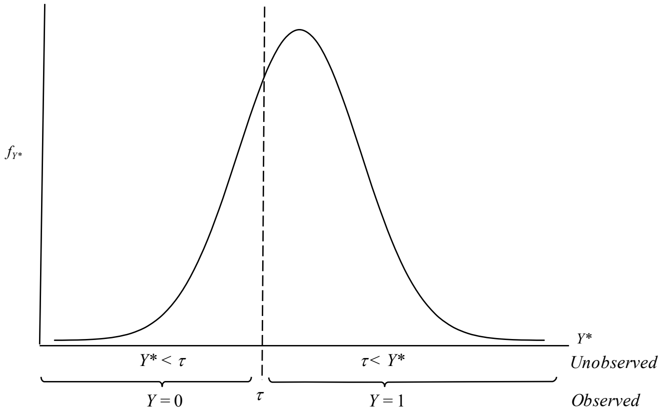
$$\Phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X} \exp\left(-\frac{1}{2}Z^2\right) dZ$$

Z is a standardized value, p is the mathematical constant, and $\exp$ is the exponential function. The figure below illustrates the concept, using Y as the observed score, Y^*, and τ (tau) as the threshold.

[1] Probit regression was developed by Edwin Wilson and Jane Worcester (1943) before logistic regression. Logistic regression, which was developed by Joseph Berkson (1944), was developed afterward but has become the much more dominant form of regression for binary outcomes.

[2] This use of the word "latent" is different from that used in structural equation modeling (SEM). In SEM, latent variables are estimated from several measures ("indicators") to account for measurement error in prediction.

[3] Although the Y^* distribution is assumed to be normal, we do not have to posit that there is a continuum underlying the observed binary Y in order to use probit regression. It is enough to simply think of a *propensity* for Y to equal 1 given some value of X even if no continuum actually exists.



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 (although we are not actually standardizing Y, we are using a normal transformation of the predicted values). The next step is to 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. Because probit values are essentially standardized scores units, the normal cdf can be used in a computational spreadsheet (e.g., `=NORM.DIST(A1,0,1,TRUE)` in Excel) or in a statistical software package to obtain the predicted probability that $Y = 1$ given the obtained values of B_0 and B_1 for some particular chosen value of X.

The probit regression is related to *polychoric* correlations, which does not require designation of an explanatory and response variable (i.e., a symmetric measure of association). 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).[4] The concept of Y^* is the same as that invoked to conceptualize probit analysis, where the polychoric correlation represents the correlation between two Y^* variables. 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 (canonical link) and error distributions (variance) 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 $\pi / \sqrt{3} \approx 1.81$, where π is the mathematical constant pi; Long, 1997, p. 48).

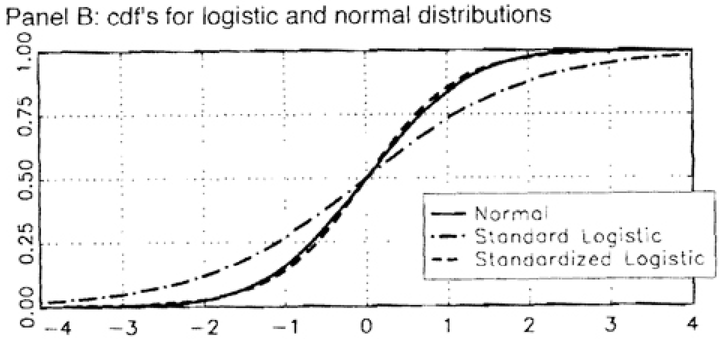


Figure 3.3. Normal and Logistic Distributions

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

[4] 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. Note that these are different from the special cases of the regular Pearson correlation, such as phi, point-biserial, or Spearman's rho correlations, because, with polychoric correlations corrections are being made to their magnitude.

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 1.81 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.

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 (logit link and log-log link), I also use μ, for the expected probability, $E(\hat{p})$ as is common in the generalized linear modeling literature. 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: $\ln\left(\frac{\mu}{1-\mu}\right)$

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

Log-log link: $\ln\left[-\ln\left(1-\mu\right)\right]$

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

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

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 forthcoming *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`), SAS (`PROC GENMOD`), and `glm` in R, allow users to specify link functions and distributions for a particular analytic circumstance.

References and Suggested Reading

Agresti, A. (2013). *Categorical data analysis (Third Edition).* New York: John Wiley & Sons.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association, 39*(227), 357-365.

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.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods, 7*, 19-40.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical. Society Series A, 135,* 370-384.

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.

Wilson, E. B., & Worcester, J. (1943). The determination of LD 50 and its sampling error in bio-assay. *Proceedings of the National Academy of Sciences*, 29(2), 79-85.

[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. The standardized logistic variance, which is approximately 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.

Probit Example

I retested the multiple logistic model with probit just to compare the results of the two types of models. First, here are the logistic results again for comparison.

Logistic Results for Comparison

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 [a]	w1sex	-.978	.214	20.824	1	.000	.376	.247	.572
	w1activ	-.041	.048	.721	1	.396	.960	.874	1.055
	w1cesd9	.035	.022	2.550	1	.110	1.036	.992	1.082
	w1neg	.068	.186	.132	1	.716	1.070	.743	1.542
	Constant	-1.199	.206	33.922	1	.000	.301		

a. Variable(s) entered on step 1: w1sex, w1activ, w1cesd9, w1neg.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	602.480 [a]	.033	.055

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

SPSS

```
plum w1hheart with w1sex w1activ w1cesd9 w1neg
/link = probit
/print= parameter summary.
```

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	503.849			
Final	480.495	23.355	4	.000

Link function: Probit.

Pseudo R-Square

Cox and Snell	.033
Nagelkerke	.056
McFadden	.037

Link function: Probit.

Parameter Estimates

							95% Confidence Interval	
		Estimate	Std. Error	Wald	df	Sig.	Lower Bound	Upper Bound
Threshold	[w1hheart = .00]	.729	.118	37.897	1	.000	.497	.960
Location	w1sex	-.548	.119	21.124	1	.000	-.782	-.314
	w1activ	-.024	.027	.781	1	.377	-.076	.029
	w1cesd9	.020	.013	2.631	1	.105	-.004	.045
	w1neg	.032	.106	.094	1	.759	-.175	.240

Link function: Probit.

In SPSS, obtain the standardized coefficients by first standardizing the predictors (make sure the same *N* is used), `descriptives vars=w1sex(zsex) w1activ(zactiv) w1cesd9(zcesd9) w1neg(zneg) /save`. and then use the same `plum` command but with `zsex`, `zactiv`, `zcesd9`, and `zneg` as the predictors. **Only use the coefficients from this run and ignore the significance tests**.

R

```
> probmod2 <- glm(w1hheart ~ w1sex + w1activ + w1cesd9 + w1neg, family=binomial(link="probit"), data=mydata)
> summary(probmod2)

Call:
glm(formula = w1hheart ~ w1sex + w1activ + w1cesd9 + w1neg, family = binomial(link = "probit"), 
    data = d)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.72853    0.11834  -6.156 7.46e-10 ***
w1sex       -0.54823    0.11928  -4.596 4.30e-06 ***
w1activ     -0.02360    0.02670  -0.884    0.377    
w1cesd9      0.02044    0.01260   1.622    0.105    
w1neg        0.03245    0.10588   0.306    0.759    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.72  on 691  degrees of freedom
Residual deviance: 602.36  on 687  degrees of freedom
AIC: 612.36

Number of Fisher Scoring iterations: 4
```

Notice that SPSS and R give the intercept (threshold) with a different sign. You can use the `reghelper` package and the function `beta` to obtain the standardized solution. (The standardized coefficients from R also could be obtained in the same way I obtained them in SPSS, by first standardizing the predictor variables, rerunning the analysis, and using only the coefficients, now standardized coefficients.) Here are the standardized results:

```
> #ignore significance tests--report the tests from the unstandardized output
> #see Long p. 70 for discussion of standardized probit
> library(reghelper)
> beta(probmod2, x = TRUE, y = FALSE)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.99954    0.05881 -16.997  < 2e-16 ***
w1sex.z     -0.26725    0.05815  -4.596  4.3e-06 ***
w1activ.z   -0.05283    0.05976  -0.884    0.377
w1cesd9.z    0.09659    0.05955   1.622    0.105
w1neg.z      0.01836    0.05990   0.306    0.759
```

Write-up

I have not included a write-up example for the probit analysis. It would proceed exactly as with the logistic analysis except there are no odds ratios to report. I would encourage you to add the standardized coefficients, however.