A Quick Guide to Logistic Regression for Qualitative, Binary Dependent Variables

1 Introduction

Often, in the social sciences, we are interested in social phenomena that are neither quantitative nor continuous (e.g. level of earnings, level of trade), but instead, are *qualitative* or *discrete* in nature. For example, the event of interest may or may not occur (e.g. war/no war, impose sanctions/no sanctions), a person makes one choice but not the other (e.g. vote/do not vote, drive to campus/take the bus). More exactly, the above are examples of what we call *dichotomous* dependent variables, for two outcomes are defined each time. For mathematical reasons, they are usually represented by *binary* values, where 1 = the event occurs ("presence"/"success"), and 0 = the event does not occur ("absence"/"failure").

Coding a dichotomous variable in binary form is mathematically appealing, as it produces conditional means, or expected values of Y, that are equal to the proportion of ones out of the total number of ones and zeros, given some X:[1]

$$E(Y|X=x) = \frac{\sum Y_x}{\sum N_{Y_x}} = Proportion(Y=1|X=x) \tag{1}$$

As proportions are often interchangeable with probabilities, the conditional means are then interpreted as probabilities of the event occurring, given some X.

1.1 Why Not Linear Regression?

If we are interested in modeling the probability of the binary outcome as a function of some explanatory variables, then linear regression analysis may be immediately appealing. However, it turns out that linear regression—such as Ordinary Least Squares, but note that some of the points made subsequently apply to the classical linear model in general—is inappropriate for regressing binary dependent variables (on a set of explanatory variables) for the following reasons: nonsensical predictions; non-linearity in variable effects; non-normal and heteroskedastic errors.

1.1.1 Nonsensical Predicted Values

By "nonsensical predictions," I mean that predicted values, or probabilities, from linear regression may violate the laws of probability. In order to understand this, we must examine the functional form of the classical linear model:

$$Y = f(X,\beta) + \epsilon \tag{2}$$

$$\hat{Y} = X\hat{\beta} \tag{3}$$

$$E(Y|X) = X\beta \tag{4}$$

[1] Mathematical notation is greatly abused throughput. Please forgive.

Equation (2) states that Y is specified as a function, f, of variables X and parameters β. We rarely see this formulation, because f is almost always a linear additive *identity link function*, i.e. $f(X,\beta) = X\beta$, resulting in the linear additive formulations for predicted values and conditional expectations in (3) and (4) respectively (and we see these almost all the time). However, this purely linear functional form $E(Y|X) = P(Y = 1|X) = X\beta$ is unbounded from negative to positive infinity, and may thus generate predicted probabilities less than zero or greater than one, which is nonsensical because the laws of probability are violated. This flaw is considered the most serious detraction against linear regression of a binary DV.

1.1.2 Nonlinear Covariate Effects

[I dare not claim to be capable enough to discuss this problem. I humbly suggest you consult Fred C. Pampel's *Logistic Regression: A Primer*, pages 5–8 (2000, Quantitative Applications in the Social Sciences, no. 132, Sage Publications) for a most excellent explanation.]

1.1.3 Non-Normal Residuals

As Y is a binary random variable with only two possible values, it follows that only two values of residuals exist for each X, which means that the disturbances are distributed Bernoulli instead of normal, violating the assumption of normally distributed disturbances for each X. The predicted probability in the linear model for each X equals $X\hat{\beta}$, hence the residuals take either of the following two values:

$$\begin{aligned} &1 - X\hat{\beta}, \quad \text{when } Y = 1 \\ &0 - X\hat{\beta}, \quad \text{when } Y = 0 \end{aligned}$$

The distribution of the residuals can never be normal when the distribution has only two values.

1.1.4 Heteroskedastic Residuals

The disturbance term also violates the assumption of homoskedasticity because the residuals vary with the value of X. As they are distributed Bernoulli:

$$Var(\hat{\epsilon}) = (X\hat{\beta})(1 - X\hat{\beta}) \tag{5}$$

If the variances are equal for all X, they would have no relationship to X. But Equation (5) shows the opposite—X values influence the size of the disturbances, which are greatest when $X\hat{\beta} = .5$, and become smaller as $X\hat{\beta}$ moves from the midpoint. As a result of heteroskedasticity, linear regression of binary DVs is inefficient, produces incorrect standard errors, and thus, incorrect estimates of statistical significance.

Admittedly, heteroskedasticity can be corrected with robust standard errors, Generalized Least Squares (GLS) or Weighted Least Squares (WLS), but these techniques do not resolve the problems of nonsensical predictions, nonlinear covariate effects, and non-normal residuals.

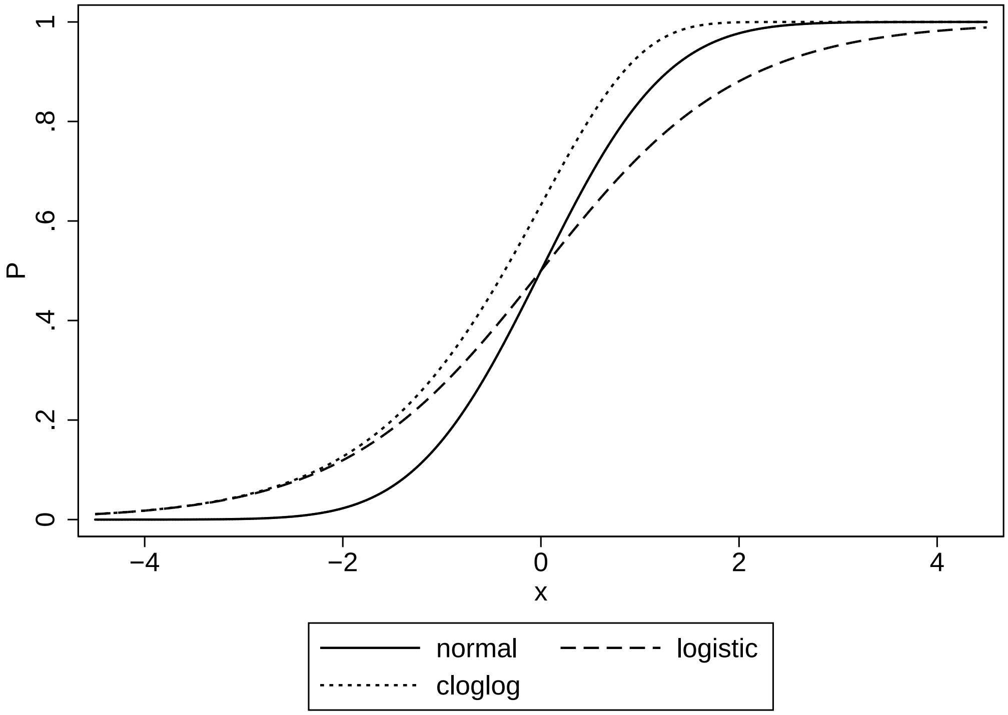

Figure 1: Cumulative Distribution Functions

1.2 An Appropriate Functional Transformation

So, we want a statistical estimator for binary DVs that produces meaningful predicted probabilities, by ensuring that the predicted values do not break the laws of probability. From a "transformational" standpoint, this means that we want to find some function F that will transform $X\beta$ from continuous and unbounded, into the appropriate form for $E(Y)$:

$$E(Y) = F(X\beta) = P(Y = 1)$$

Cumulative Distribution Functions (CDFs) fit the bill, because they are bounded between zero and one. Examples (see Figure 1)[2] include the standard normal, standard logistic, complementary log-log (note that the cloglog function is asymmetrical), and Cauchy. If the standard normal CDF, Φ, is used, we get probit regression. Using the standard logistic CDF gives us logistic regression, which is arguably more advantageous than probit as it offers more ways to interpret the parameter estimates, and thus, ability to discuss the substantive impact of explanatory variables. I discuss some of the ways of interpreting logistic regression coefficients in the next few subsections.

[2] Figure 1 can be created in Stata8 with the following command: `graph twoway function y=norm(x), range(-4.5 4.5) yvarlab(normal) || function y=invlogit(x), range(-4.5 4.5) yvarlab(logistic) || function y=invcloglog(x), range(-4.5 4.5) yvarlab(cloglog) ytitle(P)`

1.3 Predicted Probability

Let π represent $P(Y = 1)$, which is the probability of observing the outcome of interest. The formula for calculating π is:

$$\pi = \Lambda(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{6}$$

where Λ denotes the logistic CDF $\frac{e^{(\cdot)}}{1+e^{(\cdot)}}$. You will encounter another functional form for π. (Or maybe you have already encountered that alternative functional form.) Here is proof that the two functional forms are equivalent:

$$\begin{aligned}
\frac{e^{X\beta}}{1 + e^{X\beta}} &= \frac{e^{X\beta}}{1} \times \frac{1}{1 + e^{X\beta}} \\
&= \frac{1}{e^{-(X\beta)}} \times \frac{1}{1 + e^{X\beta}} \\
&= \frac{1}{e^{-(X\beta)} + \left(e^{-(X\beta)}\right)\left(e^{X\beta}\right)} \\
&= \frac{1}{e^{-(X\beta)} + e^{0}} \\
&= \frac{1}{1 + e^{-(X\beta)}}
\end{aligned}$$

1.4 Odds

Odds are defined as the ratio of the probability of observing the outcome to the probability of not observing it. Letting $\pi = P(Y = 1)$:

$$\begin{aligned}
\frac{\pi}{1-\pi} &= \frac{e^{X\beta}}{1 + e^{X\beta}} \div \left(1 - \frac{e^{X\beta}}{1 + e^{X\beta}}\right) \\
&= \frac{e^{X\beta}}{1 + e^{X\beta}} \times \frac{1}{1 - \frac{e^{X\beta}}{1+e^{X\beta}}} \\
&= \frac{e^{X\beta}}{1 + e^{X\beta} - \frac{e^{X\beta}}{1+e^{X\beta}} - \frac{e^{2(X\beta)}}{1+e^{X\beta}}} \\
&= \frac{e^{X\beta}}{1 + e^{X\beta} - \frac{e^{X\beta} - e^{2(X\beta)}}{1+e^{X\beta}}} \\
&= \frac{e^{X\beta}}{1 + \frac{e^{X\beta}(1+e^{X\beta}) - e^{X\beta} - e^{2(X\beta)}}{1+e^{X\beta}}} \\
&= \frac{e^{X\beta}}{1 + \frac{e^{X\beta} + e^{2(X\beta)} - e^{X\beta} - e^{2(X\beta)}}{1+e^{X\beta}}} \\
&= \frac{e^{X\beta}}{1 + \frac{0}{1+e^{X\beta}}} \\
&= e^{X\beta}
\end{aligned} \tag{7}$$

From Equation (7) we can see that the odds is calculated by taking the exponent of $X\beta$. An odds of 1 means that the chances of observing the outcome to not observing it are 50%/50%; an odds of 3 means that the chances are 75%/25%; an odds of $\frac{1}{3}$ means that the chances are 25%/75%, etc.

1.4.1 The Logit Transformation

Note that taking the natural logarithm of the odds produces:

$$\ln(Odds) = \ln\left(\frac{\pi}{1-\pi}\right) = \ln(e^{X\beta}) = X\beta \tag{8}$$

This is known as the *logit transformation*, which is a linear transformation of the nonlinear odds (back) to $X\beta$, which is linear additive. Thus, in binary logistic regression, $X\beta$ is also called the *logit*, or the (natural) *log odds*.

1.5 Odds Ratio

An exponentiated coefficient in binary logistic regression has the interpretation: The ratio of the odds if the corresponding variable is increased by 1, to the odds without an increase in the corresponding variable. Or, equivalently,

$$\frac{P(Y=1|x+1)/P(Y=0|x+1)}{P(Y=1|x)/P(Y=0|x)} \tag{9}$$

Here is proof, using scalar notation, that the *odds ratio*, as given in Equation (9), is equal to an exponentiated coefficient:

$$\begin{aligned}\frac{e^{a+b(x+1)}}{e^{a+bx}} &= \frac{e^{a+bx+b}}{e^{a+bx}} \\ &= e^{(a+bx+b)-(a+bx)} \\ &= e^{b} \qquad (10)\end{aligned}$$

For example, consider a hypothetical model logit vote_gore female party_ID. If the exponentiated coefficient for female = 1.5, it means that the odds of voting for Al Gore are 50% greater when female = 1 than when female = 0, with all other variables held constant. If the exponentiated coefficient for party_ID is 0.5, then the odds of voting for Gore halve as identification shifts one point to the right, and they halve at every (one point) shift/increase (again with all other variables held constant).

Do not confuse the odds with the odds ratio. (You will be surprised. Even some statistics textbooks err by doing so.) The odds is the ratio of the probability of experiencing the event to the probability of not observing it. The odds ratio is the *ratio of two odds*; the first odds with a one unit increase in one of the variables (with the remaining variable values held constant), while in the second odds all variable values are held constant.

1.6 Relative Risk

The relative risk, or risk ratio, is defined as: the ratio of the probability of observing the outcome, given a one unit increase in a variable with all other variables held constant, to the probability of observing the outcome with all variables held constant.

$$\frac{P(Y=1|x+1)}{P(Y=1|x)} \tag{11}$$

Again, be careful and do not be confused between relative risk, odds, and odds ratio. The advantage of logistic regression over probit regression is that all these alternative forms of interpreting results are available, but this also means that more care must be taken when one calculates and reports them.