CSCU Cornell Statistical Consulting Unit

Introduction to Logistic Regression

Karen Grace-Martin

Researchers are often interested in setting up a model to analyze the relationship between some predictors (i.e., independent variables) and a response (i.e., dependent variable). Linear regression is One assumption of linear models is that the residual errors follow a normal distribution. This assumption fails when the response variable is categorical, so an ordinary linear model is not appropriate. This newsletter presents a regression model for a response variable that is dichotomous having two categories. Examples are common: whether a plant lives or dies, whether a survey respondent agrees or disagrees with a statement, or whether an at-risk child graduates or drops out from high school.

In ordinary linear regression, the response variable (Y) is a linear function of the coefficients (B0, B1, etc.) that correspond to the predictor variables (X1, X2, etc.). A typical model would look like:

$$Y = B0 + B1 * X1 + B2 * X2 + B3 * X3 + \ldots + E$$

For a dichotomous response variable, we could set up a similar linear model to predict individuals' category memberships if numerical values are used to represent the two categories. Arbitrary values of 1 and 0 are chosen for mathematical convenience. Using the first example, we would assign Y = 1 if a plant lives and Y = 0 if a plant dies.

This linear model does not work well for a few reasons. First, the response values, 0 and 1, are arbitrary, so modeling the actual values of Y is not exactly of interest. Second, it is really the probability that each individual in the population responds with 0 or 1 that we are interested in modeling. For example, we may find that plants with a high level of a fungal infection (X1) fall into the category "the plant lives" (Y) less often than those plants with low level of infection. Thus, as the level of infection rises, the probability of a plant living decreases.

Thus, we might consider modeling P, the probability, as the response variable. Again, there are problems. Although the general decrease in probability is accompanied by a general increase in infection level, we know that P, like all probabilities, can only fall within the boundaries of 0 and 1. Consequently, it is better to assume that the relationship between X1 and P is sigmoidal (S-shaped), like in Figure 1, rather than a straight line.

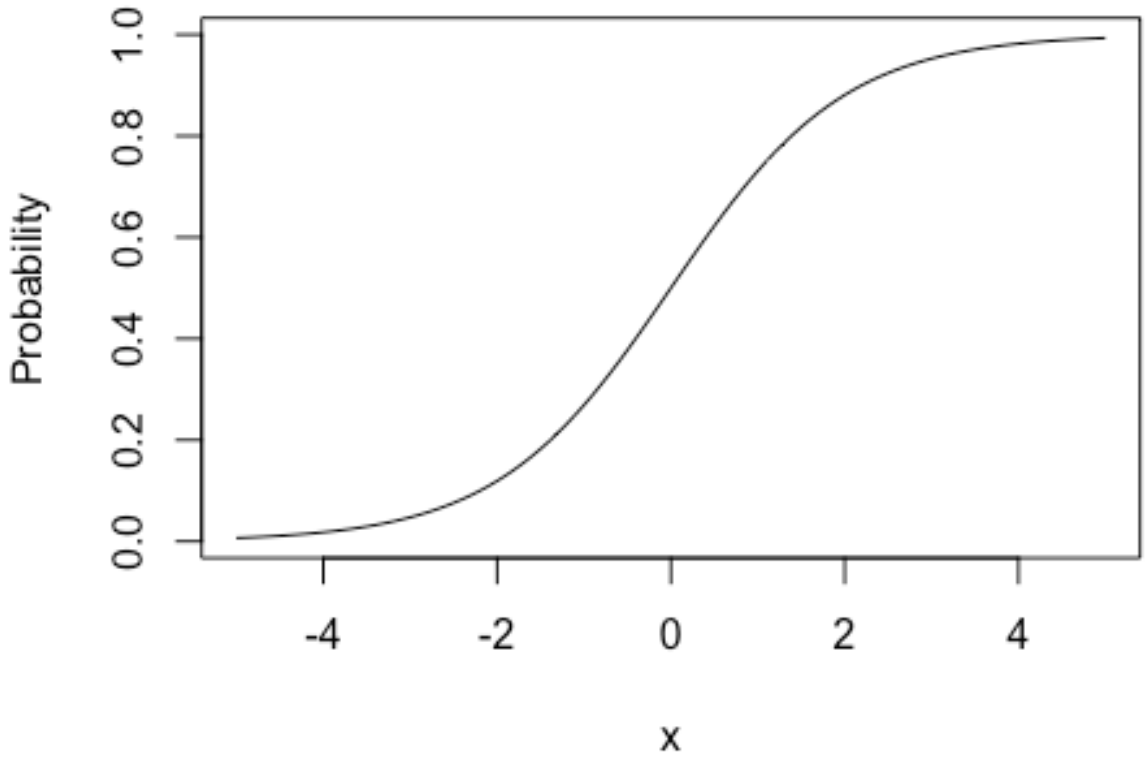

Figure 1: Graph of sigmoidal curve

It is possible, however, to find a linear relationship between X1 and a function of P. Although a number of functions work, one of the most useful is the logit function. It is the natural log of the odds that Y is equal to 1, which is simply the ratio of the probability that Y is 1 divided by the probability that Y is 0. The relationship between the logit of P and P itself is sigmoidal in shape. The regression equation that results is:

$$\ln\left(\frac{P}{1-P}\right) = B0 + B1 * X1 + B2 * X2 + \cdots$$

Although the left side of this equation looks intimidating, this way of expressing the probability results in the right side of the equation being linear and looking familiar to us. This helps us understand the meaning of the regression coefficients. The coefficients can easily be transformed so that their interpretation makes sense.

The logistic regression equation can be extended beyond the case of a dichotomous response variable to the cases of ordered categories and polytymous categories (more than two categories). Upcoming newsletters will introduce these models and discuss how to interpret coefficients in logistic regression models.

Created February 2001. Last updated April 2022.