Logistic Regression

Overview: Logistic and OLS Regression Compared

Logistic regression is an approach to prediction, like Ordinary Least Squares (OLS) regression. However, with logistic regression, the researcher is predicting a dichotomous outcome. This situation poses problems for the assumptions of OLS that the error variances (residuals) are normally distributed. Instead, they are more likely to follow a logistic distribution. When using the logistic distribution, we need to make an algebraic conversion to arrive at our usual linear regression equation (which we have written as $Y = B_0 + B_1X + e$).

With logistic regression, there is no standardized solution printed. And to make things more complicated, the unstandardized solution does not have the same straight-forward interpretation as it does with OLS regression.

One other difference between OLS and logistic regression is that there is no R^2 to gauge the variance accounted for in the overall model (at least not one that has been agreed upon by statisticians). Instead, a chi-square test is used to indicate how well the logistic regression model fits the data.

Probability that Y = 1

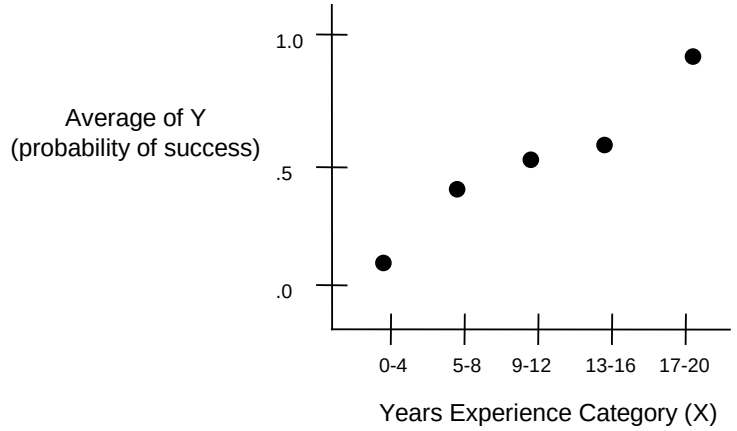
Because the dependent variable is not a continuous one, the goal of logistic regression is a bit different, because we are predicting the likelihood that *Y* is equal to 1 (rather than 0) given certain values of *X*. That is, if *X* and *Y* have a positive linear relationship, the probability that a person will have a score of *Y* = 1 will increase as values of *X* increase. So, we are stuck with thinking about predicting probabilities rather than the scores of dependent variable.

For example, we might try to predict whether or not small businesses will succeed or fail based on the number of years of experience the owner has in the field prior to starting the business. We presume that those people who have been selling widgets for many years who open their own widget business will be more likely to succeed. That means that as *X* (the number of years of experience) increases, the probability that *Y* will be equal to 1 (success in the new widget business) will tend to increase. If we take a hypothetical example, in which there were 50 small businesses studied and the owners have a range of years of experience from 0 to 20 years, we could represent this tendency to increase the probability that *Y*=1 with a graph. To illustrate this, it is convenient to break years of experience up into categories (i.e., 0-4, 5-8, 9-12, 13-16, 17-20), but logistic regression does not require this.

If we compute the mean score on Y (averaging the 0s and 1s) for each category of years of experience, we will get something like:

Yrs Exp	Average	Probability that Y=1
0-4	.17	.17
5-8	.40	.40
9-12	.50	.50
13-16	.56	.56
17-20	.96	.96

If we graph this, it looks like the following:



Notice an S-shaped curve. This is typical when we are plotting the average (or expected) values of *Y* by different values of *X* whenever there is a positive association between *X* and *Y*. As *X* increases, the probability that *Y*=1 increases. In other words, when the owner has more years of experience, a larger percentage of businesses in that category succeed. A perfect relationship represents a perfectly curved S rather than a straight line, as was the case in OLS regression. So, to model this relationship we need some fancy algebra that accounts for the bends in the curve.

The Logistic Equation

In logistic regression, a complex formula is required to convert back and forth from the logistic equation to the OLS-type equation. The logistic formulas are stated in terms of the probability that *Y* = 1, which is referred to as $\hat{p}$. The probability that Y is 0 is 1 - $\hat{p}$.

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X$$

The *ln* symbol refers to a natural logarithm and $B_0 + B_1X$ is our familiar equation for the regression line.

P can be computed from the regression equation also. So, if we know the regression equation, we could, theoretically, calculate the expected probability that *Y* = 1 for a given value of *X*.

$$\hat{p} = \frac{\exp\left(B_0 + B_1 X\right)}{1+\exp\left(B_0 + B_1 x\right)} = \frac{e^{B_0 + B_1 x}}{1+e^{B_0 + B_1 x}}$$

exp is the exponent function, sometimes written as *e.* So, the equation on the right is just the same thing but replacing *exp* with *e*. Sorry for the confusion, but *e* here is not the residual. You can always tell when *e* stands for *exp* if you see that there is a superscripted value with the *e*, suggesting that *e* is raised to some power.

Natural Logarithms and the Exponent Function. *exp*, the exponential function, and *ln*, the natural logarithm are opposites. The exponential function involves the constant with the value of 2.71828182845904 (roughly 2.72). When we take the exponential function of a number, we take 2.72

raised to the power of the number. So, exp(3) equals 2.72 cubed or $(2.72)^3 = 20.09$. The natural logarithm is the opposite of the *exp* function. If we take ln(20.09), we get the number 3. These are common mathematical functions on many calculators.

Interpretation of Coefficients. Because of these complicated algebraic translations, our regression coefficients are not as easy to interpret. Our old maxim that *b* represents "the change in Y with one unit change in X" is no longer applicable. Instead, we have to translate using the exponent function. And, as it turns out, when we do that we have a type of "coefficient" that is pretty useful. This coefficient is called the *odds ratio*.

Odds Ratio. The odds ratio is equal to *exp*(*B*), or sometimes written e^B. So, if we take the exponent constant (about 2.72) and raise it to the power of *B*, we get the odds ratio. For example, if the printout indicates the regression slope is .75, the odds ratio is approximately 2.12 (because $\exp(.75) = 2.72^{.75} = 2.12$). This means that the probability that *Y* equals 1 is twice as likely (2.12 times to be exact) as the value of X is increased one unit. An odds ratio of .5 indicates that *Y*=1 is half as likely with an increase of X by one unit (so there is a negative relationship between X and Y). An odds ratio of 1.0 indicates there is no relationship between X and Y.

This odds ratio terminology is perhaps easiest to understand when we are dealing with a special case in which both X and Y are dichotomous. When they are both dichotomous, the odds ratio is the probability that Y is 1 when X is 1 *compared* to the probability that Y is 1 when X is 0. Some authors use the Greek symbol ψ (psi, pronounced like "sci" in science) to refer to the odds ratio, and others use OR or O.R. To get *b* from the odds ratio, just take the log of the odds ratio, *ln*(ψ). For this reason, the slope is sometimes called the "log odds".

Using a chi-square-like 2 X 2 table, one can compute the odds ratio, ψ, using this formula:

$$\hat{\psi} = \frac{success\,vs.\,failure\,when\,X=1}{success\,vs.\,failure\,when\,X=0} = \frac{X=1\,when\,Y=1\,/\,X=1\,when\,Y=0}{X=0\,when\,Y=1\,/\,X=0\,when\,Y=0} = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

Here, π refers to the probability that *Y* = 1. So, $\pi(1)$ is the when *Y*=1 and *X*=1, and $\pi(0)$ is when *Y*=1 when *X*=0.[1] However, we can also use odds ratios and logistic regression when the predictor is continuous. In that case, computing it by hand is too difficult, at least in our privileged era of computers.

Model Fit

Deviance. With logistic regression, instead of R^2 as the statistic for overall fit of the model, we have deviance instead. Remember, when we studied chi-square analyses, chi-square was said to be a measure of "goodness of fit" of the observed and the expected values. We use chi-square as a measure of model fit here in a similar way. It is the fit of the observed values (*Y*) to the expected values ($\hat{Y}$). The bigger the difference (or "deviance") of the observed values from the expected values, the poorer the fit of the model. So, we want a small deviance if possible. As we add more variables to the equation the deviance should get smaller, indicating an improvement in fit.

[1] For a table labeled so that *(X=1 when Y=1*) is *a*, *(X=1 when Y=0)* is *b*, *(X=0 when Y=1*) as *c*, and *(X=0 when Y=0)* as *d*, then the odds ratio has the following short-cut equivalent formulae: *(a/c)/(b/d) = (a*d)/(b*c) = (a/b)/(c/d)*. A related concept, *relative risk* or *risk ratio*, can be distinguished from the odds ratio. The relative risk in health research is the risk of disease relative to exposure and is computed using marginal frequencies: *[a/(a+b)]/[c/(c+d)]*. With rare conditions, relative risk and odds ratio are very similar. Some areas of research (e.g., clinical trials) prefer to use the relative risk measure and for some it is considered more intuitive. The odds ratio, however, is probably more widely used and has more direct connection to logistic regression.

Maximum Likelihood. Instead of finding the best fitting line by minimizing the squared residuals, as we did with OLS regression, we use a different approach with logistic— Maximum Likelihood (ML). ML is a way of finding the smallest possible deviance between the observed and predicted values (kind of like finding the best fitting line) using calculus (derivatives specifically). With ML, the computer uses different "iterations" in which it tries different solutions until it gets the smallest possible deviance or best fit. Once it has found the best solution, it provides a final value for the deviance, which is usually referred to as "negative two log likelihood" (shown as "-2 Log Likelihood" in SPSS). The deviance statistic is called –2LL by Cohen et al. (2003) and Pedazur and *D* by some other authors (e.g., Hosmer and Lemeshow, 1989), and it can be thought of as a chi-square value.

The likelihood ratio test, G: A chi-square difference test using the "null" or constant-only model. Instead of using the deviance (*–2LL*) to judge the overall fit of a model, however, another statistic is usually used that compares the fit of the model with and without the predictor(s). This is similar to the change in R^2 when another variable has been added to the equation. But here, we expect the deviance to decrease, because the degree of error in prediction decreases as we add another variable. To do this, we compare the deviance with just the intercept (*–2LL*$_{null}$ referring to *–2LL* of the *constant-only* model*)* to the deviance when the new predictor or predictors have been added (*–2LL*$_k$ referring to *–2LL* of the model that has *k* number of predictors). The difference between these two deviance values is often referred to as *G* for goodness of fit (important note: *G* is referred to as "chi-square" in SPSS printouts).

$$G = \chi^2 = D\left(\textit{ for the model without the variable}\right) - D\left(\textit{ for the model with the variable}\right)$$

or, using the Cohen et al. notation,

$$\begin{aligned} G = \chi^2 &= D_{null} - D_k \\ &= -2LL_{null} - (-2LL_k) \end{aligned}$$

where D_{null} is the deviance for the constant only model an D_k is the deviance for the model containing *k* number of predictors. An equivalent formula sometimes presented in textbooks is:

$$G = \chi^2 = -2\ln\left(\frac{L_{null}}{L_k}\right)$$

where the ratio of the ML values is taken before taking the log and multiplying by –2. This gives rise to the term “likelihood ratio test” to describe *G*.

One can look up the significance of this test in a chi-square table using *df* equal to the number of predictors added to the model (but the test is also provided in the printout). The chi-square values reported in the SPSS printout compare the –2LL for the model tested to the -2LLfor a model with just the constant (i.e., no predictors), but one could use the difference in deviance values to compare any two nested logistic models.