Logistic Regression

Logistic regression involves a prediction of a binary outcome. Ordinary least squares (OLS) regression assumes a continuous dependent variable Y that is distributed approximately normally in the population. Because a binary response variable will not be normally distributed and because the form of the relationship to a binary variable will tend to be nonlinear, we need to consider a different type of model.

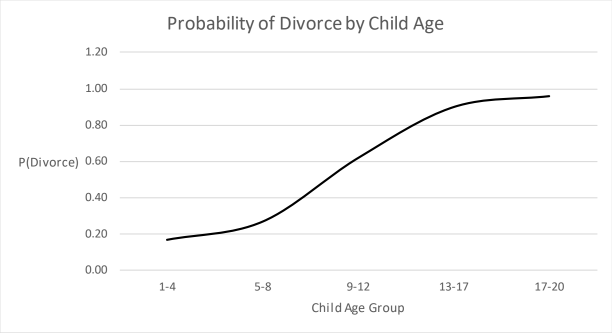
Predicting the Probability that $Y = 1$

For a binary response variable, we can frame the prediction equation in terms of the probability of a discrete event occurring. Usual coding of the response variable is 0 and 1, with the event of interest (e.g., "yes" response, occurrence of an aggressive behavior, or heart attack), so that, 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.

For example, we might try to predict whether or not a couple is divorced based on the age of their youngest child. Does the probability of divorce ($Y = 1$) increase as the youngest child's age (X) increases? If we take a hypothetical example, in which there were 50 couples studied and the children have a range of ages from 0 to 20 years, we could represent this tendency to increase the probability that $Y = 1$ with a graph, grouping child ages into four-year intervals for the purposes of illustration. Assuming codes of 0 and 1 for Y, the average value in each four-year period is the same as the estimated probability of divorce for that age group.

Child Age	Average	Probability of
	$E(Y\|X)$	Divorce ($Y = 1$)
1-4	0.17	0.17
5-8	0.27	0.27
9-12	0.62	0.62
13-17	0.90	0.90
17-20	0.96	0.96

The average value within each age group is the expected value for the response at a given value of X, which, with a binary variable, is a conditional probability. Graphing these values, we get



Notice the 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, assuming a normal and equal distributions for X at each value of Y. As X increases, the probability that $Y = 1$ increases, but not at a consistent rate across values of X. In other words, when children are older, an increasing larger percentage of parents in that child age category divorce, with the increase in divorce probability more dramatic for the middle child age groups.

The Logistic Equation

The S-shaped curve is approximated well by a natural log transformation of the probabilities. In logistic regression, a complex formula is required to convert back and forth from the logistic equation to the OLS-type equation. The logistic equation is stated in terms of the probability that $Y = 1$, which is $\hat{p}$ (the caret symbol ^ is used by the text to underscore that the probability is a sample estimate), and the probability that $Y = 0$, which is $1 - \hat{p}$.[1]

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

The natural log transformation of the probabilities is called the *logit transformation*. The right hand side of the equation, B_1X+B_0, is the familiar equation for the regression line. The left hand side of the equation, $\ln\left(\hat{p}/1-\hat{p}\right)$, referred to as the *logit*, stands in for the predicted value of Y (the observed values are not transformed). So, the predicted regression line is curved line, because of the log function. With estimates of the intercept, B_0, and the slope B_1, $\hat{p}$ can be computed from the equation using the complementary function for the logarithm, e. Given a particular value of X, we can calculate the expected probability that Y = 1.

$$\hat{p} = \frac{e^{(B_1X+B_0)}}{1+e^{(B_1X+B_0)}}$$

Because the intercept is the value of Y when X equals 0, the estimate of the probability of $Y = 1$ when $X = 0$ is $\hat{p} = e^{B_0} / \left(1 + e^{B_0}\right)$.

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.

Regression Coefficients and Odds Ratios

Because of the log transformation, our old maxim that β represents "the change in Y with one unit change in X" is no longer applicable. The exponential transformations of the regression coefficient, β, gives us the *odds ratio*, however, which has an understandable interpretation of the increase in odds for *each unit increase in X*. For illustration purposes, I used grouped ages, in which case, a unit increase would be from one group to the next. Nearly always, we would rather use a more continuous version of age, so a unit increase might be a year. If the odds ratio was 1.22, we would expect approximately a 22% increase in the probability of divorce with each increment in child age. We need to be a little careful about such interpretations, and realize that we are talking about an average percentage increase over all of the range of X. Look back at table of divorce probabilities and the S-shaped figure above. We do not see the same increment in the probability of divorce from the first child age category to the second as we do between the second and the third.

For the special case in which both X and Y are 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.

$$OR = e^{\beta_1} = \frac{n_{21}/n_{22}}{n_{11}/n_{12}} = \frac{\hat{P}_{21}/\hat{P}_{22}}{\hat{P}_{11}/\hat{P}_{12}} = \frac{\hat{P}_{11}\hat{P}_{22}}{\hat{P}_{21}\hat{P}_{12}}$$

Caution is needed in interpreting odds ratios less than 1 (negative relationship) in terms of percentages, because 1/1.22 = .82, where you might be tempted to (incorrectly) interpret the value as indicating an 18% decrease in the probability of divorce instead of more accurately, a 22% decrease. The farther away from

[1] Note that the Snijders and Bosker (2012) notation in Chapter 17 is a little inconsistent, but I use P-hat for proportions as they do but focus on level-1 equation notation, using β instead of γ, to illustrate basic logistic principles. I've left off the subscripts for the regression coefficients, assuming constant non-varying coefficients for now.

1.0, the bigger this discrepancy is (e.g., 1/.4 = 2.5, suggesting a 150% decrease rather than a 60% decrease).

Odds ratios require some careful interpretation generally because they are essentially in an unstandardized metric. Consider using age as measured by year instead of category in the divorce example. We would expect a smaller percentage increase in the probability that $Y = 1$ for each unit increase in X if X is per year rather per four-year interval increase. If a predictor is measured on a fine-grained scale, such as dollars for annual income, each increment is miniscule and would not the percentage increase in the event to be very large, even if there is a strong magnitude of the relationship between the income and the event.

Standardized Coefficients

To address the magnitude interpretation problem with odds ratios, the X variable is sometimes standardized to obtain the odds increase for each standard deviation increase in X, which is sometimes referred to as a *partially standardized* coefficient. Fully standardized coefficients for logistic regression also can be computed, although their meaning is less straightforward than in ordinary least squares regression and there is no universally agreed upon approach. Because software programs do not implement any of them, researchers rarely if ever consider reporting them. A standardized coefficient would have the advantage of interpretation for understanding the relative contribution of each predictor. One can simply calculate the standard deviations of X and Y and standardize the logistic regression coefficient using their ratio as is done in ordinary least squares regression, $\beta_1 = B_1(s_x/s_y)$. Menard (2010; 2011) suggests using the standard deviation of the logit, sd^2_{logit}, and the R^2 value as defined for ordinary least squares regression.[2]

$$\beta_1 = \frac{sd_x B_1}{\sqrt{sd^2_{\text{logit}} / R^2}}$$

Significance Tests and Confidence Intervals for β and Odds Ratios

The significance of the regression coefficient (that $B \neq 0$ in the population) can be tested with the Wald ratio,

$$Wald\ \chi^2 = \left(\frac{B}{SE_B}\right)^2$$

The test may be expressed as a z-test in some software, where $Wald\ z = \sqrt{Wald\ \chi^2}$. The standard error computation is complex and is derived from the maximum likelihood estimation iterative process. Although the Wald test is the most commonly employed, because it is printed for each coefficient in all software packages, it does not perform optimally in all circumstances. For smaller samples, tends to be too conservative (i.e., Type II errors are more likely—true relationships are not found to be significant) for large coefficients (Hauck & Donner, 1977; Jennings, 1986). Confidence intervals can also be constructed

$$B \pm (1.96) SE_B$$

where 1.96 is the z critical value for the normal distribution when α = .05 two-tailed. If the confidence interval includes zero, then the coefficient is nonsignificant. Odds ratios may also be presented with confidence limits, in which case, an interval that includes 1.0 is nonsignificant.

Model Fit

Maximum likelihood estimation is used to compute logistic model estimates. The iterative process finds the minimal discrepancy between the observed response, Y, and the predicted response, $\hat{Y}$ (see the handout "Maximum Likelihood Estimation"). The resulting summary measure of this discrepancy is the -2 loglikelihood or -2LL, known as the *deviance* (McCullagh & Nelder, 1989). The larger the deviance, the

[2] Menard (2011, Appendix) describes the details for the computer steps required to compute the variance of the standard deviation of the logit (sd^2_{logit}) and standardized coefficients. In multilevel logistic, the variance of the logit could be obtained, but one of the R-square values might have to be substituted for R-square from the OLS model.

larger the discrepancy between the observed and expected values. A smaller deviance represents a better fit. The concept is similar to the mean square error (MSE) in ANOVA or regression. Smaller MSE indicates better fit and better prediction. As we add more predictors to the equation, the deviance should get smaller, indicating an improvement in fit. The deviance for the model with one or more predictors is compared to a model without any predictors, called the *null model* or the *constant only* model, which is a model with just the intercept. The now familiar likelihood ratio test is used to compare the deviances of the two models (the null model, L_0 and the full model, L_1).

$$\begin{aligned} G^2 &= Deviance_0 - Deviance_1 \\ &= -2\ln\left(\frac{L_0}{L_1}\right) = \left[-2\ln\left(L_0\right)\right] - \left[-2\ln\left(L_1\right)\right] \end{aligned}$$

The estimated value of G^2 is distributed as a chi-squared value with df equal to the number of predictors added to the model. The loglikelihoods from any two models can be compared as long as the same number of cases are used and one of the models has a subset of the predictors used in the other model. The special case of the likelihood ratio test in which just one variable is added to the model gives a likelihood ratio test of the significance of a single predictor—the same hypothesis tested by the Wald ratio described above. A third alternative, the *score* test (or Lagrange multiplier test) is also based on partial derivatives of the likelihood function evaluated at α (i.e., the intercept β_0). The score test is not printed in most software packages for individual parameters and is not reported very often by researchers. The Wald, likelihood ratio, and score tests will usually give a very similar result, and are in fact asymptotically equivalent (Cox & Hinkley, 1972), but the likelihood ratio and score test tend to perform better in many situations (e.g., Hauck & Donner, 1977). The Wald test assumes a symmetric confidence interval whereas the likelihood ratio does not.

To assess overall fit of the model, the Pearson χ^2 and likelihood ratio test (G^2) do not perform well (McCullagh 1985), especially when data are *sparse* (expected values in a multi-way contingency table are small, < 5). A test that is commonly reported in software output is one developed by Hosmer and Lemeshow (Lemeshow, 2000), but this test does not perform well under many conditions. There have been a variety of proposed alternatives (see Allison, 2014 for an excellent summary), most of which are preferable, but not ideal. Moreover, most are not available from most software programs. You will also hear about several absolute fit indices such as the Akaike information criteria (AIC) Bayesian information criteria (BIC), which can be useful for comparing models (lower values indicate better fit) but are not informative without a comparison.

Pseudo R^2 Measures

There is not an easily defined R^2 with logistic regression that can be used to quantify the variance accounted for in the response variable, but there are some propsed pseudo-R^2 values based on improvement in fit (reduction in deviance) when one or more variables are added to the model. The most common are the Cox and Snell (Cox & Snell, 1989; Cragg & Uhler, 1970; Maddala,1983) and Nagelkerke (1991) pseudo R^2 values. You may also see those proposed by McFadden (1974) and Tjur (2009), among others (see Allison, 2014, for a review). Each have values that theoretically range between 0 and 1.

$$R^2_{Cox\,\&\,Snell} = 1 - \left[\frac{-2L_0}{-2L_1}\right]^{2/n} \qquad R^2_{Nagelkerke} = \frac{1 - \left[\dfrac{-2L_0}{-2L_1}\right]^{2/n}}{1 - \left(-2L_0\right)^{2/n}}$$

Generalized Linear Models

The transformation used in logistic regression is a transformation of the predicted scores of Y or the predicted line, not the observed Y values. 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). 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.

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 *Regression Models for Count Data* handout in my multiple regression class). 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.