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.

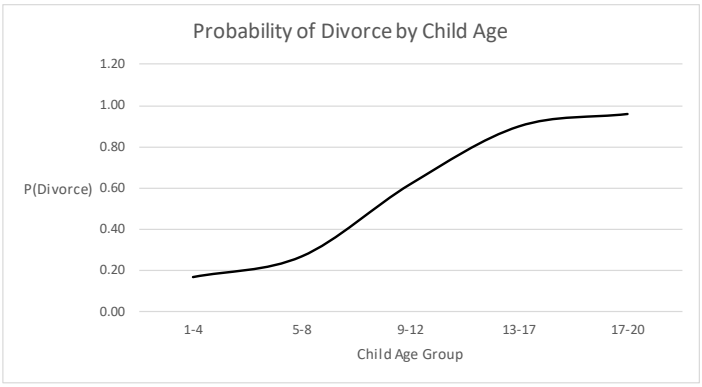
Predicting the Probability that $Y = 1$

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 $E(Y\|X)$	Probability of 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 increasingly 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 $1 - \hat{p}$.

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

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, $\ln(\hat{p}/1-\hat{p})$, 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 $\hat{p} = e^{B_0}/(1+e^{B_0})$.

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 B_1 represents "the change in Y with one unit change in X" is no longer applicable. The exponential transformations of the regression coefficient, B_1, using e^B or $\exp(B1)$ gives us the *odds ratio*, however, which has a more 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 1.53, we would expect approximately a 53% 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.

When X is Binary
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, much like the percent comparisons we made with the chi-square analysis of contingency tables. [1]

$$OR = e^{B_1} = \frac{n_{21}/n_{22}}{n_{11}/n_{12}} = \frac{\hat{p}_{21}/\hat{p}_{22}}{\hat{p}_{11}/\hat{p}_{12}} = \frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{21}\hat{p}_{12}}$$

	Y=1	Y=0
$X = 1$	n_{11}	n_{12}
$X = 0$	n_{21}	n_{22}

Sometimes abbreviated with cell labels *a, b, c, d,* where $OR = e^{B_1} = \frac{ad}{cb}$

	Y=1	Y=0
= 1	*a*	*b*
$X = 0$	*c*	*d*

Caution is needed in interpreting odds ratios less than 1 (negative relationship) in terms of percentages, because 1/.82 = 1.22, 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 1.0, 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 one year in age or dollars for annual income, each increment is miniscule and the percentage increase in the odds that the event occurs would not be very large, even if there is a strong magnitude of the relationship between the predictor and the event.

[1] For tables 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)*.

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.

Relative Risk

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. The relative risk can be obtained from the odds ratio, because $RR = OR\left[\left(1-\hat{p}_{1+}\right)/\left(1-\hat{p}_{2+}\right)\right]$ if the marginal frequencies ($\hat{p}_{1+}$ for the first row and $\hat{p}_{2+}$ for the second row) are known.

Multiple Logistic Regression

Like ordinary least squares regression, a logistic regression model can include two or more predictors. The coefficients and the odds ratios then represent the effect of each independent variable controlling for all of the other independent variables in the model and each coefficient can be tested for significance. Any combination of binary and continuous predictors is possible. For nominal predictors with more than two categories, a set of g -1 dummy variables need to be constructed and entered together to capture the differences among the g groups.

As with multiple regression, we may want also to assess the overall fit of the model and to know whether all of the predictors, taken together, account for a significant amount of variance in the dependent variable. Logistic regression, however, has no simple-to-define multiple R^2, and the assessment of model fit is more complicated than with OLS regression. Overall model fit will be discussed in the handout "Multiple Logistic Regression and Model Fit."

[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.

References and Further Reading

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, *81*(394), 471-476.

Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, *72*(360a), 851-853.

Hosmer D.W. and S. Lemeshow (1980). A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics A10:1043-1069.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. New York: Wiley.

Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications, second edition*. Sage Publications.

Menard, S. (2011). Standards for standardized logistic regression coefficients. *Social Forces, 89*, 1409-1428.

Newsom, J.T., Rook, K.S., Nishishiba, M., Sorkin, D., & Mahan, T.L. (2005). Understanding the relative importance of positive and negative social exchanges: Examining specific domains and appraisals. *Journals of Gerontology: Psychological Sciences, 60B*, P304-P312.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychology and Measurement, 1, 385-401.

Santor, D. A. & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9*, 233–243.

Sorkin, D. H., & Rook, K. S. (2004). Interpersonal control strivings and vulnerability to negative social exchanges in later life. *Psychology and Aging, 19*, 555–564.