Logistic Regression

Michael P. LaValley, PhD

Like contingency table analyses and χ^2 tests, logistic regression allows the analysis of dichotomous or binary outcomes with 2 mutually exclusive levels.[1] However, logistic regression permits the use of continuous or categorical predictors and provides the ability to adjust for multiple predictors. This makes logistic regression especially useful for analysis of observational data when adjustment is needed to reduce the potential bias resulting from differences in the groups being compared.[2]

Use of standard linear regression for a 2-level outcome can produce very unsatisfactory results. Predicted values for some covariate values are likely to be either above the upper level (usually 1) or below the lower level of the outcome (usually 0). In addition, the validity of linear regression depends on the variability of the outcome being the same for all values of the predictors. This assumption of constant variability does not match the behavior of a 2-level outcome. So, linear regression is not adequate for such data, and logistic regression has been developed to fill this gap.

Some recent examples of use of logistic regression in *Circulation* include the assessment of gender as a predictor of operative mortality after coronary artery bypass grafting surgery,[3] an evaluation of the relationship between the TaqIB genotype and risk of cardiovascular disease in a meta-analysis,[4] and an examination of the relationship between lipoprotein abnormalities and the incidence of diabetes.[5]

The Logistic Regression Model

The logistic regression model has its basis in the odds of a 2-level outcome of interest. For simplicity, I assume that we have designated one of the outcome levels the event of interest and in the following text will simply call it the event. The odds of the event is the ratio of the probability of the event happening divided by the probability of the event not happening. Odds often are used for gambling, and "even odds" (odds=1) correspond to the event happening half the time. This would be the case for rolling an even number on a single die. The odds for rolling a number <5 would be 2 because rolling a number <5 is twice as likely as rolling a 5 or 6. Symmetry in the odds is found by taking the reciprocal, and the odds of rolling at least a 5 would be 0.5 (=1/2).

The logistic regression model takes the natural logarithm of the odds as a regression function of the predictors. With 1 predictor, X, this takes the form $\ln[\text{odds}(Y=1)]=\beta_0+\beta_1 X$, where ln stands for the natural logarithm, Y is the outcome and Y=1 when the event happens (versus Y=0 when it does not), β_0 is the intercept term, and β_1 represents the regression coefficient, the change in the logarithm of the odds of the event with a 1-unit change in the predictor X. The difference in the logarithms of 2 values is equal to the logarithm of the ratio of the 2 values, so by taking the exponential of β_1, we obtain the ratio of the odds (the odds ratio) corresponding to a 1-unit change in X.

Odds ratios often are used in the analysis of 2-by-2 contingency tables[6] and case-control studies.[7] The odds ratio is sometimes confused with the relative risk, which is the ratio of probabilities rather than odds. Only when the probability of the event is very low can the odds ratio be considered a good approximation to the relative risk.[2] The odds ratio is more extreme than the relative risk, which leads to exaggeration of the effect of a predictor when it is misinterpreted as a relative risk.[8] In many settings, the relative risk is preferred over the odds ratio because it addresses the more readily understood probability of the event rather than its odds.[9] However, logistic regression results are typically presented by odds ratios because these are the natural estimates from the model and attempts to transform these to relative risks can distort the results.[10]

A useful way to think of the odds ratio is that 100 times the odds ratio minus 1, ie, 100×(odds ratio−1), gives the percent change in the odds of the event corresponding to a 1-unit increase in X. If this value is negative, then the odds of the event decrease with increasing values of X; if positive, the odds increase. This percentage change is the same for any 1-unit increase in X because of the assumed linearity between X and the logarithm of the odds in the regression model above. For some continuous predictors, this assumption may not match the data,[11] in which case careful checking of the model results is required. For example, if the logarithm of the odds against the predictor X has a U shape (both low and high values have large odds of the outcome relative to the intermediate values) and the model assumes a linear (straight line) pattern, then goodness-of-fit checking should show that the model and the data are not compatible. In such a case, splitting the predictor values into categories and using dummy variables to code for the categories may improve the fit.[1] Other methods such as splines also may be used to lessen the assumption of linearity.[12]

From the Department of Biostatistics, Boston University School of Public Health, Boston, Mass.
Correspondence to Dr Michael P. LaValley, Department of Biostatistics, Boston University School of Public Health, 715 Albany St, Crosstown Center Room 322, Boston, MA 02118. E-mail mlava@bu.edu

(*Circulation*. 2008;117:2395-2399.)

Circulation is available at http://circ.ahajournals.org

DOI: 10.1161/CIRCULATIONAHA.106.682658

Table 1. Unadjusted and Adjusted Odds Ratios for Development of Angina

	Unadjusted			Adjusted		
Predictor	Odds Ratio	95% CI	*P*	Odds Ratio	95% CI	*P*
Cholesterol (1 SD)	1.412	(1.297, 1.537)	<0.001	1.404	(1.284–1.535)	<0.001
Sex				1.415	(1.173–1.705)	<0.001
Current smoking				1.035	(0.854–1.255)	0.728
Diabetes				1.437	(0.891–2.320)	0.138
Age (10 y)				1.088	(0.973–1.216)	0.139
Body mass index (1 SD)				1.299	(1.190–1.419)	<0.001
Heart rate (1 SD)				0.867	(0.788–0.953)	0.0031

Odds ratios, 95% CIs, and probability values for predictors of angina in the Framingham data. Columns 2 through 4 present results from the unadjusted model; columns 5 through 7 show results from the adjusted model. The respective SDs for cholesterol, body mass index, and heart rate are 44.622 mg/dL, 4.077 kg/m^2, and 12.033 bpm.

When adjusted values are needed, more predictors can be added to the right side of the regression equation above, along with corresponding regression coefficients (β). In this case, the odds ratio value for X would be adjusted for the other predictors in the model. The equation above, 100×(odds ratio−1), would then be interpreted as the percent change in the odds corresponding to a 1-unit increase in X while holding all other predictors fixed. The selection of appropriate predictors to reduce confounding and to improve the precision of estimates is done similarly for logistic regression and for linear regression; guidelines can be found in many statistical textbooks.[1,2,12]

Unlike linear regression, there is no formula for the estimates of β for logistic regression. Finding the best estimates requires repeatedly improving approximate estimates until stability is reached. This is done easily on a computer, and there are many statistical software packages that perform logistic regression, but it makes logistic regression less understandable and more of a "black box" approach for many researchers.

Angina in the Framingham Heart Study

To illustrate the use of logistic regression, I use data from the Framingham Heart Study[13] that are available for teaching purposes from the National Heart, Lung, and Blood Institute (http://www.nhlbi.nih.gov/resources/deca/teaching.htm). These data include subjects at the 1956 Framingham examination, considered to be the baseline, with 24 years of follow-up. Here, I analyze the event of development of new angina pectoris during the follow-up. Subjects with prevalent angina at the 1956 examination are excluded from the data, and only measures from the 1956 examination are used as predictors. Not all subjects have complete 24-year follow-up because some died or left the study before 1980. Use of survival analysis methods to account for varying length of follow-up[14] would be appropriate for a more definitive study of these data.

The predictor of main interest in my analysis is the measure of serum total cholesterol (mg/dL), and I consider adjusting for the sex of the subject, current smoking (yes or no), presence of diabetes (yes or no), age (years), body mass index (kg/m^2), and ventricular heart rate (bpm). All of the analyses were done with SAS version 9.1 (SAS Institute Inc, Cary, NC).

After those with prevalent angina are removed, 4287 subjects remain, and 578 subjects (13.5%) developed new angina during the follow-up. At the 1956 examination, 56.8% of subjects were women, 49.5% were current smokers, and 2.9% had diabetes. The mean total cholesterol was 236.7 mg/dL (limits, 107 to 696 mg/dL), mean age was 49.6 years (limits, 32 to 70 years), mean body mass index was 25.8 kg/m^2 (limits, 15.5 to 56.8 kg/m^2), and mean heart rate was 75.9 bpm (limits, 44 to 143 bpm).

Table 1 gives the unadjusted and adjusted odds ratios for a difference of 1 SD (44.622 mg/dL) of cholesterol on the occurrence of new angina during the follow-up. In the unadjusted model, cholesterol is the only predictor; in the adjusted model, sex, current smoking, presence of diabetes, age, body mass index, and heart rate also are included. In the unadjusted model, there is a 41.2% increase in the odds of angina with each 1-SD increase in total cholesterol, and there is a 40.4% increase in the adjusted model. Often, there is greater discrepancy between adjusted and unadjusted estimates. So, in these data, there is little confounding of the effect of cholesterol as a result of the other predictors in the adjusted model. From the adjusted model, the odds of angina are increased 42% for men compared with women, and increased body mass index and decreased heart rate increase the odds of angina. The effects of current smoking, the presence of diabetes, and age are not larger than could be due to chance in these data ($P>0.05$).

In a data set with fewer cases of angina, the confidence interval for the adjusted result could be wider owing to increasing the variability of the estimates when more predictors are used than the data would support. A rule of thumb for stability of the estimates from logistic regression is to have at least 10 events (or nonevents, whichever is rarer in the data) per predictor in the model—more precisely, per degree of freedom used in the model.[15] Because there are about 83 cases of angina for each predictor in the adjusted model, the results are quite stable.

Goodness of Fit

One aspect of the results of logistic regression that is not described in the preceding section is how well the model

Table 2. Hosmer and Lemeshow Test Results for Unadjusted and Adjusted Logistic Regression Models

	Unadjusted Model		Adjusted Model	
Predicted Probability Ranking Groups	Observed Angina Cases, n	Expected Angina Cases, n	Observed Angina Cases, n	Expected Angina Cases, n
1 (Lowest)	26	34.4	22	23.2
2	26	40.1	31	31.6
3	53	43.8	41	38.4
4	55	49.2	37	44.4
5	60	52.0	56	50.4
6	57	57.4	63	56.0
7	69	61.3	57	62.8
8	65	65.6	70	71.2
9	72	75.4	83	83.1
10 (Highest)	90	93.8	112	111.1
χ^2	13.6		4.0	
P		0.094		0.854

Hosmer and Lemeshow test results for the prediction of angina in the Framingham data. Columns 2 and 3 show the observed and expected numbers of angina cases by group for the unadjusted model. Columns 4 and 5 show the observed and expected numbers of angina cases by group for the adjusted model. χ^2 Test statistics (on 8 *df*) and the probability values are shown for each model.

agrees with the observed data. This is called the goodness of fit of the model. The odds ratio values given above describe the model as it is applied to the data. If the model and the data are not in good agreement, then these odds ratios are not very meaningful.[16] Several authors have pointed out that although goodness of fit is crucial for the assessment of the validity of logistic regression results in medical research, it often is not included in published articles.[16–18]

Goodness of fit is usually evaluated in 2 parts. The first step is to generate global measures of how well the model fits the whole set of observations; the second step is to evaluate individual observations to see whether any are problematic for the regression model.[1] Some global measures of goodness of fit include R^2 measures for logistic regression; the c statistic, a measure of how well the model can be used to discriminate subjects having the event from subjects not having the event; and a test of model calibration developed by Hosmer and Lemeshow.[19] The second part of evaluating goodness of fit is focused on looking for outliers and influence points and may be useful for seeing whether linearity in the model is reasonable.

The R^2 measures for logistic regression mimic the widely used R^2 measure from linear regression, which gives the fraction of the variability in the outcome that is explained by the model. However, logistic regression R^2 does not have such intuitive explanation, and values tend to be close to 0 even for models that fit well. Because there is an upper bound for the basic logistic regression R^2, a rescaled R^2 is usually also presented showing the fraction of the upper bound that is attained. In the logistic regressions predicting angina, the model containing only cholesterol as a predictor had an R^2 of 0.015 with a rescaled R^2 of 0.0275. The model containing 7 predictors had an R^2 of 0.0304 and a rescaled R^2 of 0.0555. The adjusted model has larger R^2 values, but it is difficult to judge whether the difference is large enough to be important.

The c statistic measures how well the model can discriminate between observations at different levels of the outcome. It is the same as the area under the receiver-operating characteristic curve,[20] formed by taking the predicted values from the regression model as a diagnostic test for the event in the data. The minimum value of c is 0.5; the maximum is 1.0. In their textbook, Hosmer and Lemeshow[1] consider c values of 0.7 to 0.8 to show acceptable discrimination, values of 0.8 to 0.9 to indicate excellent discrimination, and values of $\geq$0.9 to show outstanding discrimination (page 162). The c statistic value is 0.603 in the unadjusted model for angina and 0.643 in the adjusted model, both below the threshold for acceptable discrimination.

The Hosmer and Lemeshow test evaluates whether the logistic regression model is well calibrated so that probability predictions from the model reflect the occurrence of events in the data. Obtaining a significant result on the test would indicate that the model is not well calibrated, so the fit is not good. For this test, subjects are grouped by their percentile of predicted probability of having the event according to the model: group 1 has subjects with predicted probabilities in the 1st to 10th percentiles, group 2 has subjects with predicted probabilities in the 11th to 20th percentiles, and so on. If the observed and expected numbers of events are very different in any group, then the model is judged not to fit. Observed and expected values for the groups in the unadjusted and adjusted models for angina are shown in Table 2. The unadjusted model has a borderline-significant (P=0.094) test result, indicating possible problems with the model fit. In the adjusted model, the test finds less evidence of lack of fit (P=0.854). Inspection of Table 2 shows that the adjusted model has much better agreement between observed and expected numbers of angina events, especially for groups with low percentages of expected events, ie, in subjects with relatively low cholesterol.

Problematic points are those that are either outliers, data values for which the observed value and the model prediction are in poor agreement, or influence points, observations with

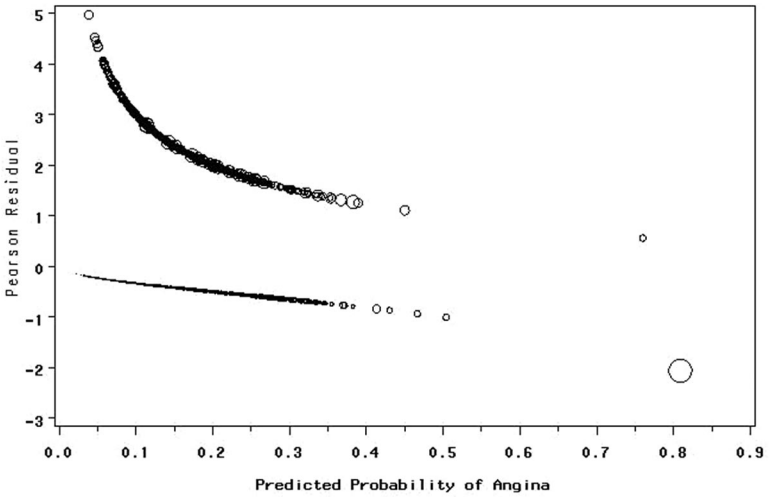

Figure. Residual plot from the adjusted model for angina in the Framingham data. The horizontal axis shows the predicted probability of angina; vertical axis, the value of the Pearson residual. The size of the plotted circle is proportional to the influence of an observation.

an unexpectedly large impact on model results. Checking for problematic observations is done by plotting residuals against predicted values, the model estimate of the probability that a subject will have the event.[21] Outliers are observations with large residuals, and in logistic regression, several residuals have been developed. Here, I use the relatively simple Pearson residual, which is the difference between the observed and expected outcomes for an observation divided by the square root of the variability of the expected outcome. Logistic regression residual plots look different from those from linear regression because the residuals fall on 2 curves, 1 for each outcome level. Pearson residuals >3 and <-3 would be considered potential problems, although for large data sets we should expect some values beyond those limits. There also are several measures of influence for logistic regression. Here, I use the logistic regression version of Cook's distance, which provides a measure of how much the model estimates change when each point is removed. Neither outliers nor influence points should be discarded automatically, but having knowledge of their presence can be used for targeted data checking and cleaning, or sensitivity analyses.

The Figure is a residual plot for the adjusted model. The horizontal axis shows the predicted probability of angina for each observation; the vertical axis shows the Pearson residual. The size of the plotted circle is proportional to the Cook's distance for the observation. The higher curve is of subjects who developed angina, and the lower curve is of subjects who did not. Because the number of subjects who developed angina is smaller, their observations are generally more influential, and their circles tend to be larger. From the Figure, we can identify several possible problems. First, there are 2 observations with predicted probabilities of angina between 0.75 and 0.80. These come from 2 subjects with unusually high cholesterol values (600 and 696 mg/dL). The subject with 696 mg/dL did not develop angina, making a rather poor fit to the model and the most influential observation in these data, shown by having the largest circle. There are also subjects who developed angina despite having a very low predicted probability in the model. The low predicted probabilities for these subjects were primarily due to low cholesterol values. The mismatch between the observed angina rates and low predicted probability of angina in the regression model for these subjects creates large residuals, and these are the points in the upper left region of the Figure. A substantial number of these subjects have residual values >3 and might be considered outliers.

So, although we cannot reject that the adjusted model fits the data according to the Hosmer and Lemeshow test, the R^2 and c values are still rather low. In addition, the Figure makes it clear that there are some subjects with low cholesterol who develop angina and are not well fit by the model. There are also some subjects with very high cholesterol who may have excessive influence on the model estimates. As a sensitivity analysis, we might want to remove subjects with cholesterol of $\geq$600 mg/dL and see if the model results change substantially. We also might consider adding more predictors or allowing a nonlinear effect of cholesterol to see if we can better predict angina for subjects with low cholesterol levels.

Extensions to the Logistic Regression Model

Here, I have considered only outcomes with 2 levels, but there are extensions to the logistic regression model that allow analysis of outcomes with $\geq$3 ordered levels such as no pain, moderate pain, or severe pain. Such data often are analyzed with proportional odds logistic regression,[22] although other models also are possible.[23,24] Multinomial logistic regression may be used if the outcome consists of $\geq$3 unordered categories.[1] The standard form of logistic regression presented here also presumes that observations are independent. This would not be the case for longitudinal or clustered data, and analyzing such data as independent could give misleading conclusions.[25] Methods such as generalized estimating equations[26] or random-effects models[27] can be used for such data. Finally, survival analysis methods[14] provide an extension for studies in which subjects have been followed up for events across extended and varying follow-up times.

Disclosures

None.

References

1. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York, NY: John Wiley & Sons, Inc; 2000.

2. Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. Oxford, UK: Blackwell Science Ltd; 2003.
3. Blankstein R, Ward RP, Arnsdorf M, Jones B, Lou YB, Pine M. Female gender is an independent predictor of operative mortality after coronary artery bypass graft surgery: contemporary analysis of 31 Midwestern hospitals. *Circulation*. 2005;112(suppl):I-323–I-327.
4. Boekholdt SM, Sacks FM, Jukema JW, Shepherd J, Freeman DJ, McMahon AD, Cambien F, Nicaud V, de Grooth GJ, Talmud PJ, Humphries SE, Miller GJ, Eiriksdottir G, Gudnason V, Kauma H, Kakko S, Savolainen MJ, Arca M, Montali A, Liu S, Lanz HJ, Zwinderman AH, Kuivenhoven JA, Kastelein JJ. Cholesteryl ester transfer protein TaqIB variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment: individual patient meta-analysis of 13,677 subjects. *Circulation*. 2005;111:278–287.
5. Festa A, Williams K, Hanley AJ, Otvos JD, Goff DC, Wagenknecht LE, Haffner SM. Nuclear magnetic resonance lipoprotein abnormalities in prediabetic subjects in the Insulin Resistance Atherosclerosis Study. *Circulation*. 2005;111:3465–3472.
6. Bland JM, Altman DG. Statistics notes: the odds ratio. *BMJ*. 2000; 320:1468.
7. Breslow NE, Day NE. Statistical methods in cancer research, volume I: the analysis of case-control studies. *IARC Sci Publ*. 1980:5–338.
8. Holcomb WL Jr, Chaiworapongsa T, Luke DA, Burgdorf KD. An odd measure of risk: use and misuse of the odds ratio. *Obstet Gynecol*. 2001;98:685–688.
9. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? *BMJ*. 1998;316:989–991.
10. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157:940–943.
11. Lee J. An insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence. *Int J Epidemiol*. 1986;15:22–29.
12. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer-Verlag; 2001.
13. Fox CS, Pencina MJ, Meigs JB, Vasan RS, Levitzky YS, D'Agostino RB Sr. Trends in the incidence of type 2 diabetes mellitus from the 1970s to the 1990s: the Framingham Heart Study. *Circulation*. 2006;113:2914–2918.
14. Hosmer DW, Lemeshow S. *Applied Survival Analysis*. New York, NY: John Wiley & Sons; 1999.
15. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373–1379.
16. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health*. 1991; 81:1630–1635.
17. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol*. 2001;54:979–985.
18. Bender R, Grouven U. Logistic regression models used in medical research are poorly presented. *BMJ*. 1996;313:628.
19. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Commun Stat*. 1980;A10:1043–1069.
20. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press; 2003.
21. Friendly M. *Visualizing Categorical Data*. Cary, NC: SAS Institute Inc; 2000.
22. Bender R, Grouven U. Ordinal logistic regression in medical research. *J R Coll Physicians Lond*. 1997;31:546–551.
23. Harrell FE Jr, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, Muhe L, Gatchalian S, Eichenwald HF. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants: WHO/ARI Young Infant Multicentre Study Group. *Stat Med*. 1998;17:909–944.
24. Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. *J Clin Epidemiol*. 1997;50:45–55.
25. Cannon MJ, Warner L, Taddei JA, Kleinbaum DG. What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. *Stat Med*. 2001;20:1461–1467.
26. Lipsitz SR, Kim K, Zhao L. Analysis of repeated categorical data using generalized estimating equations. *Stat Med*. 1994;13:1149–1163.
27. Twisk JWR. *Applied Longitudinal Data Analysis for Epidemiology*. Cambridge, UK: Cambridge University Press; 2003.

KEY WORDS: angina ■ epidemiology ■ risk factors ■ statistics