More on Model Fit and Significance of Predictors with Logistic Regression

Overall Model Fit

The most common assessment of overall model fit in logistic regression is the goodness-of-fit test (G), which is simply the chi-square difference between the null model (i.e., with the constant only) and the model containing one or more predictors. This is one use of the likelihood ratio test between two nested models (referrred to as "chi-square" in the SPSS logistic output). It is an assessment of the improvement of fit between the predicted and observed values on Y by adding the predictor(s) to the model.

In some cases, the traditional goodness-of-fit test (G or the likelihood ratio test) may not be the best assessment of model fit. Some simulations suggest that the deviance statistic is not distributed as chi-square when the data are *sparse*. The term "sparse" refers to a circumstance in which there are few observed values (and therefore few expected values) in the cells formed by crossing all of the values of all of the predictors. The Hosmer-Lemeshow fit test is designed to correct for this (when there are continuous predictors). Note that they do not recommend the use of this test when there is a small n (less than 400; Hosmer & Lemeshow, 2000). The Hosmer-Lemeshow test is performed by dividing the predicted probabilities into deciles (10 groups based on percentile ranks) and then computing a Pearson chi-square that compares the predicted to the observed frequencies (in a 2 X 10 table). Lower values (and nonsignificance) indicate a good fit to the data and, therefore, good overall model fit.

R^2 for Logistic Regression

In logistic regression, there is no true R^2 value as there is in OLS regression. However, because deviance can be thought of as a measure of how poorly the model fits (i.e., lack of fit between observed and predicted values), an analogy can be made to sum of squares residual in ordinary least squares. The proportion of *unaccounted* for variance that is reduced by adding variables to the model is the same as the proportion of variance accounted for, or R^2.

$$R^2_{\text{logistic}} = \frac{-2LL_{null} - 2LL_k}{-2LL_{null}} \qquad\qquad R^2_{OLS} = \frac{SS_{total} - SS_{residual}}{SS_{total}} = \frac{SS_{regression}}{SS_{total}}$$

Where the null model is the logistic model with just the constant and the *k* model contains all the predictors in the model.

In SPSS, there are two modified versions of this basic idea, one developed by Cox & Snell and the other developed by Nagelkerke. The Cox and Snell R-square is computed as follows:

Cox & Snell Pseudo-R^2

$$R^2 = 1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{2/n}$$

Because this R-squared value cannot reach 1.0, Nagelkerke modified it. The correction increases the Cox and Snell version to make 1.0 a possible value for R-squared.

Nagelkerke Pseudo-R^2

$$R^2 = \frac{1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{2/n}}{1 - \left(-2LL_{null}\right)^{2/n}}$$

Tests of a Single Predictor

In the case of a simple logistic regression (i.e., only a single predictor), the tests of overall fit and the tests of the predictor test the same hypothesis: is the predictor useful in predicting the outcome? The Wald test is the usual test for the significance of a single predictor (is $B_{pop} = 0$ or is $OR_{pop} = 1.0$). Thus, for simple logistic both the likelihood ratio for the full model and the Wald test for the significance of the predictor test the same hypothesis. A third alternative is the Score test (sometimes referred to as the "Lagrange Multiplier" test).

The Likelihood Ratio, Wald, and Score test of the significance of a single predictor are said to be "asymptotically" equivalent, which means that their significance values will converge with larger N. With small samples, however, they are not likely to be equal and may sometimes lead to different statistical conclusions (i.e., significance). The likelihood ratio test for a single predictor is usually recommended by logistic texts as the most powerful (although some authors have stated that neither the Wald nor the LR test are superior). Wald tests are known to have low power (higher Type II errors) and can be biased when there is insufficient data (i.e., expected frequency is too low) for each category or value of X. However, I have seen very few researchers use the likelihood ratio test for tests of individual predictors. One reason may be that the statistical packages do not provide this test for each predictor, making hand computations and multiple analyses necessary. This is inconvenient, especially for larger models. If the analysis has a large N, researchers are likely to be less concerned about the differences. There seems to be less known about the performance of the Score test and it is not currently available in SPSS.

Comments

The above approaches to calculating R-squared with logistic regression are only two of several different approaches. At this point, there does not seem to be much agreement on which approach is best, and researchers do not seem to report either very often when logistic analyses are performed. My recommendation would be to use these without considering them to be definitive values for the percentage of variance accounted for and to make some reference to their "approximate" accuracy.

References and Further Reading

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression (2nd Edition)*. New York: Wiley.

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

O'Connel, A.A. (2006). *Logistic regression models for ordinal response variables.* Thousand Oaks: Sage. QASS #146.

Summary Table of Statistical Tests in Logistic Regression

	Alternative Terms	Statistical Description	Notes
Overall Model Fit			
Deviance	D, Deviance chi-square, -2LL, -2 log likelihood	Based on minimization of the maximum likelihood function	Can be computed for any model, distributed as chi-square value
Likelihood Ratio Test	G, “Chi-square” in SPSS, LR test, nested model chi-square test	$G = \chi^2 = -2LL_{null} - (-2LL_k)$ or equivalently, $G = \chi^2 = -2\ln\left(\frac{L_{null}}{L_k}\right)$	Comparison of null or constant only model to the full model which includes the predictors. Can be used to compare any two “nested” models.
Hosmer & Lemeshow Goodness of Fit Test	none	Peason chi-square is used in a special procedure where a continuous predictor is categorized into several groups	Can provide improved estimates of fit when the sample size is large. With small samples (with n < 400, according to Hosmer & Lemeshow, 2000), its use is not recommended.
Psuedo R^2s	Cox & Snell, Nagelkerke, R_L^2	See previous logistic handout	
Predictor Significance			
Wald Chi-square	none, occasionally presented as a z-test	$\frac{B^2}{SE_B^2}$	Most commonly used test of significance of an individual predictor (B_{pop}=0), distributed as chi-square with one df
Score Test	Lagrange Multiplier test (LM)	uses first derivative of likelihood function for B=0 (the Wald is based on the second derivative).	
Likelihood Ratio Test	see above	Same computation as in the above section but the “null model” is replaced by a model with one fewer predictors. The difference in fit is then a test of a single predictor.	compares model with and without a particular predictor