Multiple Logistic Regression and Model Fit

Multiple Logistic Regression Overview

Just as in OLS regression, logistic models can include more than one predictor. The analysis options are similar to regression. The researcher can enter the predictors simultaneously or they can be entered in blocks (hierarchical). There are also exploratory options, as with a stepwise procedure. The interpretation of the results from a multiple logistic regression is similar to interpretation of the results from a multiple OLS regression. Slopes and odds ratios represent the "partial" prediction of the dependent variable. A slope for a given predictor represents the expected average change in logit of Y for each unit change in X, holding constant the effects of the other variable. Overall fit of the model with multiple predictors is more complicated than with OLS regression.

Unlike OLS regression, there is no true R-squared value and the test of the significance of a set of predictors taken together is not an F test. Instead, tests of whether a set of variables together account for a significant amount of variance is conducted with a likelihood ratio test (printed as the “chi-square test” in SPSS), sometimes called “goodness-of-fit test” or G^2, that is computed for any two models that are nested. Nested models are models in which only a subset of predictors from the full model are included. Adding predictors will improve the fit of the model. A chi-square test is not valid unless the two models compared involve one model that is a reduced form of (i.e., nested within) the other model. In particular, the two models must be based on the same set of cases.

Model Estimation and Basics of Fit

Maximum likelihood estimation is used to compute logistic model estimates.[1] This iterative process, which is the same general process we discussed in connection with loglinear models, finds the minimal discrepancy between the observed response, Y, and the predicted response, $\hat{Y}$. 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 larger the discrepancy between the observed and expected values. The concept is similar to the mean square residual (MS_{res}) in regression or mean square error (MSE) in ANOVA. Smaller MSE indicates better fit and better prediction, or, alternatively, larger MSE indicates worse fit or lack of fit. 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 (and “Block 0” in SPSS), which is a model with just the intercept. The *likelihood ratio test* (LR test) is used to compare the deviances of the two models (the null model, L_0 and the full model, L_1).[2]

$$
\begin{aligned}
G^2 &= deviance_0 - deviance_1 \\
&= -2\ln\left(\frac{L_0}{L_1}\right) = \left[-2\ln\left(L_0\right)\right] - \left[-2\ln\left(L_1\right)\right]
\end{aligned}
$$

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 deviances 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. Most commonly, the likelihood ratio test (G^2), compares the null model (i.e., with no predictors or “constant only” or Block 0) to the model containing one or more predictors and thus provides and omnibus test of all of the predictors together, similar to the F-test of the R^2 in ordinary least squares regression. The deviance (usually referred to as -2 loglikelihood or -2LL) for each model (the null and the full model) will be printed in

---

[1] See the handout “Maximum Likelihood Estimation” on the webpage for my Categorical Data Analysis class.

[2] Important note: G^2 is referred to as "chi-square" in SPSS logistic output. And $\ln$ is the natural log, so $\ln = \log$ used in some other texts. A special case of this equation is the same as the G^2 equation we examined last term in the handout “Common Ordinal Analyses: Loglinear Models and Measures of Association” which shows that, for a 2 × 2 frequency table, G^2 is a function of the observed (N_{ij}) and expected frequencies (μ_{ij}) across each of the cells.

$$G^2 = 2\sum_i^I \sum_j^J N_{ij} \log\left(\frac{N_{ij}}{\mu_{ij}}\right)$$

the output. The likelihood ratio test, which is just the difference between these two values, also will be given along with its associated significance level. In the SPSS logistic output, the likelihood ratio (G^2) is referred to simply as "chi-square". 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 other words, whether the predictors together account for a significant amount of variance in the outcome.

Special case—one added predictor. Although the likelihood ratio test is usually used for an omnibus test of all the predictors together, a special case of the likelihood ratio test is with just one variable added to the model and so gives a test of the significance of that one predictor. That is the same hypothesis tested by the Wald ratio (as z or chi-square) described earlier that was used for the test of the regression coefficient, B. A third alternative, the *score* test (or Lagrange multiplier test) is also based on partial derivatives of the likelihood function evaluated at B_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 of a single predictor 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. Although rarely seen outside of textbooks, the Wald test also can be computed for a set of variables, but the likelihood ratio is nearly always the method of testing a set of variables added to the model.

Alternative Measures of Fit

Classification Tables. Most regression procedures print a classification table in the output. The classification table is a 2 × 2 table of the observed values on the outcome (e.g., 0="no", 1="yes) and then the values predicted for the outcome by the logistic model. Then the percentage of correctly predicted values (percent of 0s and 1s) correctly predicted by the model is given (in SPSS, be sure to look at the Block 1 classification table not the Block 0 table). To compute this, some criteria for deciding what is a correct prediction is needed, and by default the program will use the probability that Y = 1 exceeding .5 as "correct." Although authors often report percent correct from the classification table as an indicator of fit, it has an inherent problem in the use of .5 as an arbitrary cutoff for correct that is influenced by the base rate value of the probability that Y = 1 (see Box 13.2.8 in the Cohen, Cohen, West, & Aiken, 2003 text). So, I tend not to use the percent correctly classified and tend to take it with a grain of salt when other researchers report it.

*Hosmer-Lemeshow Tes*t. The likelihood ratio test (G^2) does not always perform well (Hosmer & Lemeshow, 1980; McCullagh 1985; Xu, 1996), especially when 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. An alternative test developed by Hosmer and Lemeshow (1980) is commonly printed with logistic regression output. 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 × 10 table). Lower values (and nonsignificance) indicate a good fit to the data and, therefore, good overall model fit. Unfortunately, even Hosmer and Lemeshow (2013) do not recommend using their test unless the sample size is at least 400 (when sparseness may not be as much of a problem) because of insufficient power; and it has other potential problems (Allison, 2014; Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997). There are several other potential alternative fit tests, such as the standardized Pearson test or the Stukel test, which are not widely available in software packages and appear to be less often used by researchers (see Allison, 2014 for an excellent summary), some of which may also require larger sample sizes for sufficient power (Hosmer & Lemeshow, 2013).

Information Criteria. You will also hear about several absolute fit indices, such as the Akaike information criteria (AIC) or Bayesian information criteria (BIC), which can be useful for comparing models (lower values indicate better fit). (SPSS does not print several other global fit indices that are sometimes used by researchers testing logistic regression models). The AIC and BIC do not have values that are informative by themselves because they are fairly simply derived from the deviance using adjustments for sample size and number of predictors. Because the deviance itself depends on the size of the model, variances of the variables involved, and other factors, it has no possible standard of magnitude and thus neither does the

AIC or BIC (there are no statistical tests for these indices and no cutoff for what constitutes a good fit). Indices like the AIC and BIC are occasionally used, however, to try to compare non-nested models (models that do not have the same cases or where one model is not a subset of predictors from the other model). When models are nested, the likelihood ratio (difference in deviances) can be used as a statistical test (chi-square value), so there is not really a need for the AIC or BIC in that case. The AIC and BIC are perhaps the most commonly used but there are several other similar indices, such as the AICC and aBIC. The equations below show the AIC and BIC are fairly simply derived of the deviance (-2LL value), shown below with p as the number of predictors and n as the sample size.

$$AIC = -2LL + 2(p+1)$$
$$BIC = -2LL + \log(n)(p+1)$$

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 the sum of squares residual in ordinary least squares. These are called "pseudo R-square" values. 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.

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

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

There are a number of pseudo R^2 values that have been proposed using this general logic, including the Cox and Snell (Cox & Snell, 1989; Cragg & Uhler, 1970; Maddala,1983), Nagelkerke (1991), McFadden (1974), and Tjur (2009) indexes, among others (see Allison, 2014, for a review). As two common examples, consider the following:

Cox & Snell Pseudo-R^2

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

Because the Cox and Snell 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

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

At this point, there does not seem to be much agreement on which R-square approach is best (see https://statisticalhorizons.com/r2logistic for a brief discussion and references), and researchers do not seem to report any one of them as often as they should. My recommendation for any that you choose to use is that you should not use them as definitive or exact values for the percentage of variance accounted for and you should make some reference to the "approximate percentage of variance accounted for."

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?).[3] 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 regression 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 (cf. Hosmer, Hosmer, Le Cessie, & Lemshow, 1997; Xie, Pendergast, & Clarke, 2008) at least across a range of conditions, and it is not currently available in many software packages for individual predictors (although it shows up under "variables not in the equation" in SPSS).

References and Further Reading

Allison, P. D. (2014). Measures of fit for logistic regression. *SAS Global Forum, Washington, DC*.
Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
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 Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression, third edition*. New York: Wiley.
Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods, 9*, 1043-1069.
Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine, 16*, 965-980.
Long, J.S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.
McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistics in linear exponential family models. *International Statistical Review 53*, 61–67.
McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.
Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications, second edition*. Sage Publications.
O'Connel, A.A. (2006). *Logistic regression models for ordinal response variables.* Thousand Oaks: Sage. QASS #146.
Xie, X. J., Pendergast, J., & Clarke, W. (2008). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. Computational Statistics & Data Analysis, 52(5), 2703-2713.
Xu, H. (1996). Extensions of the Hosmer-Lemeshow goodness-of-fit test (Doctoral dissertation, University of Massachusetts at Amherst).

[3] Although the Wald test can theoretically be used to test multiple coefficients simultaneously (see Long, 1997, p. 90), it is generally only used in practice and by most software programs as a test of a single coefficient (given either as a *z*-test or chi-square test).

Summary Table of Statistical Tests in Logistic Regression

	Alternative Terms	Statistical Description	Notes
Overall Model Fit			
Deviance	*D*, Deviance (or 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	$G = \chi^2 = -2LL_{null} - \left(-2LL_k\right)$ 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	Pearson 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, McFadden, Tjur	See formulas on previous page.	There is not universal consensus on which is best and there are others that have been proposed. Use as a supplement to the LR and present as "approximate" proportion of variance accounted for. Be prepared to calculate someone else's favorite value.
Predictor Significance			
Wald Chi-square	None, occasionally presented as a *z*-test rather than chi-square	$\frac{B^2}{SE_B^2}$	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).	Not very commonly reported and not currently available in SPSS for predictors in the model.
Likelihood Ratio Test	See above discussion	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, but, in SPSS, tests of a single predictor's significance must be obtained through hierarchical (nested) model comparisons.