[Journal of Contemporary Research in Education](https://egrove.olemiss.edu/jcre)

Volume 2
Number 1 *August 2013*

Article 3

8-1-2013

Logistic regression Model Effectiveness: Proportional Chance Criteria and Proportional Reduction in Error

Jeffry L. White
University of Louisiana, Lafayette

Journal of Contemporary Research in Education
2(1) 4-10

Logistic Regression Model Effectiveness: Proportional Chance Criteria and Proportional Reduction in Error

Jeffry L. White
University of Louisiana, Lafayette

Abstract

The importance of classification tables in binary logistic regression analysis has not been fully recognized. This may be due to an over reliance on statistical software or lack of awareness of the value that computation of the proportional by chance accuracy criteria (PCC) and proportional reduction in error (PRE) statistic can add to binary logistic regression models. Case illustrations are used in this paper to demonstrate the usefulness of these computations. An overview of logistic regression is proffered along with a discussion of the function of case classifications and strategies in application of the PCC and PRE. It offers guidance for others interested in understanding how classification tables can be maximized to assess the predictive effectiveness and utility of binary logistic regression models.

Introduction

The use of logistic regression analysis to predict dichotomous outcomes in education is an alternative to linear regression that has gained popularity with the availability of statistical software packages (Baradwaj & Pal, 2011; Teh, Othman & Michael, 2010). Increased use of logistic regression requires that educational researchers become knowledgeable in how to accurately assess and interpret the results (Peng, Lee, & Ingersoll, 2002). While user friendly software may have contributed to the popularity, it does not preclude the use of computational techniques to garner more meaningful information. In addition to understanding the underlying assumptions of logistic regression and principles of statistical interpretation, researchers must also evaluate the accuracy and utility of their models to determine how well they work (Menard, 2002).

Statistical programs like STATA, R, SAS, and SPSS create contingency tables of the observed and predicted values of the dependent variables similar to chi square (Menard, 2002). By comparing the predicted with the observed values (George & Mallery, 2011) the probability of a particular case is classified into one of the outcomes based on the regression equation. Classification tables are created to indicate how well the model predicts the possible values of the dependent variable by indicating the percent of overall classifications, which is a key ingredient in determining the accuracy of the model (Long, 1997). While this may be sufficient in some situations, other researchers may be more interested in determining the *utility* and *predictive efficiency* of the model rather than the overall fit. This can be accomplished via the *proportional by chance accuracy criteria* (PCC) and *proportional reduction in error* (PRE) statistic.

This paper discusses the efficacy and utility the PCC and PRE bring to binary logistic regression models. Case illustrations are presented to demonstrate their application. An overview of logistic regression is proffered along with a discussion of classifying cases and how the PCC and PRE are used to determine effectiveness and utility. It illuminates how classification tables can be used to evaluate the usefulness and efficiency of binary logistic regression models.

Overview of Logistic Regression

<u>Test of Significance</u>

Binary logistic regression (LR) is a variation of linear regression in which continuous, discrete, dichotomous, or a combination of these variables are used to predict the occurrence or non-occurrence of an event (Hair, Anderson, Tatham, & Black, 2009; Pezzullo, 2004). It can be expanded to multinomial outcomes to determine the amount of explained variance and the relative importance of each of the predictors (Garson, 2004). It also permits the investigator to assess how well the model fits the data by comparing the predictions with the observed outcomes and the utility of the variables in the prediction (Pampel, 2000).

Logistic regression applies maximum likelihood estimation after transforming the dependent variable into a logit variable. A logit variable is the natural *log of the odds* of the outcome occurring or not. In this way the logistic regression estimates the probability of the occurrence of the event (Garson, 2004).

The hypothesis is that the coefficient for the logistic regression (*Bk*) is zero. It can be interpreted as the change in the log odds associated with a one-unit change in the independent variable (Stevens, 2007). If the coefficient is positive, its value will be greater than 1, indicating a one-unit increase in the independent variable. This means the odds are increased that the event will occur. If the coefficient is negative, the value of *Bk* will be less than 1, indicating a decrease in the odds that the event will take place. If the value of *Bk* is zero, the odds remain unchanged for every one-unit increase in the independent variable.

The omnibus test of statistical significance in LR is the Wald statistic. It is calculated as the squared ratio of the logistic regression to its standard error, or Wald = $(Bk/\text{S.E.})^2$. It should be noted that the Wald statistic presents problems when the absolute value of the logistic regression coefficient is large (Stevens, 2007). The estimated standard error is inflated in large coefficients and results in lowering the Wald statistic (Menard, 2002). This can result in a failure to reject the hypothesis that the coefficient is zero and lead to an erroneous conclusion, or Type II error, that the effect is not significant when it actually is (false negative).

The contribution each independent variable makes to the model can be difficult to determine when they are highly correlated (Stevens, 2007). This is due to the basic assumption that there is no linear relationship among the independent variables (Garson, 2004). For that reason, a correlation matrix of the independent variables should be inspected. If the variables are highly correlated (> .50) their impact can be assessed by the Likelihood Ratio Test. This can be done by using the *Backward LR* entry method in SPSS and examining the *Model if Term Removed* pivot table. Each predictor is tested using the hypothesis that the full model is indistinguishable when the variable is removed. The ones with the smallest *p* values contribute the most.

Goodness of Fit

In addition to testing significance, the logistic regression model assesses the goodness-of-fit of the data. The probability of the results meeting the parameter estimates is examined using the -2 times the log of the likelihood (-2LL) as a measure of how well the model fits the data (Stevens, 2007). A good model will result in a high likelihood of the observed results (small value for -2LL). If the data fits the model perfectly the likelihood will be 1, and the -2LL will be 0.

The null hypothesis for goodness of fit is that the observed likelihood does not differ from 1. To test, the value of -2LL is used with the expectation that it has a chi square distribution with n – p degrees of freedom, where n = number of cases and p = number of parameters estimates – constant (*Bo*) + *Bk* for each predictor. The chi square statistic tests the null hypothesis that the logistic regression coefficients for all the terms in the model except the constant (*Bo*) are 0, or stated otherwise, H0: $B1 = B2 = Bk = 0$. The desired outcome is that the hypothesis is not rejected and the model fits the data (Stevens, 2007).

The *Step chi square* statistic is also used to examine the goodness of fit of the model (Stevens, 2007). It is comparable to the *F* statistic in multiple regression analysis testing the null hypothesis that the coefficients for predictor variables added at each step = 0.

Statistics Analogous to R^2

The software provides several statistics that attempt to quantify the proportion of variance explained by the LR model (Norusis, 2003) or measure the strength of association (Garson, 2004). In binary cases, SPSS automatically defaults to the Cox and Snell R^2 and McFadden's in multinomial LR. The Cox and Snell (1989) statistic presents problems for interpretation because its maximum value is usually less than 1.0. Fortunately, there are other techniques similar to R^2 available to measure the strength of association, such as Menard and Nagelkerke's Pseudo R^2 statistics (Freese & Long, 2006). In the Menard (2000), values vary from 0 (indicating that the independent variables are useless in predicting the dependent variable) to 1.0 (the model accurately predicts the dependent variable). These indices are identical in the Nagelkerke (1991) statistic and Cohen's (1983) guidelines are used to measure the effect size.

Classification of Cases

To assess how well the model fits the data, the predictions of whether the event is expected to occur or not are compared with the observed outcomes (Stevens, 2007). Statistical software like SPSS and SAS include a classification table and/or histogram of *Observed Groups and Predicted Probabilities* to assess the goodness of fit. Particular attention is paid to the percent of predicted classifications that are correct for the anticipated groups and the overall percent of correct predictions. In a perfect model, 100% of the cases will be situated on the diagonal axis (Garson, 2004).

Classification Tables

In a binary logistic regression, the classification table is a 2 x 2 contingency table of the observed and predicted results. The model is used to classify each record using the computed probabilities ranging between 0 and 1 with .50 as the minimum probability (or cut value). Data records with probabilities greater than .50 are classified as 1. Those less than .50 are assigned a value of zero (0). Cases where the event is observed to occur should scale toward high probabilities. The cases where the event is not observed should scale toward low probabilities (Stevens, 2007).

To better illustrate an example, two of the four data cells in Table 1 represent correct classifications. The other incorrect cells are referred to as false negatives (observed = 0, predicted = 1) or false positives (observed = 1, predicted = 0). In Table 1 there are 99 false positives and 37 false negatives indicating the model classification was 80.9% (157/194) correct for the predicted = 0 cases and 58.6% (140/239) correct for the 140 predicted = 1 cases. The overall fit of the model yielded 68.6% correct classifications (297/433).

Table 1: Sample classification table ($n = 433$)

Classification Table[a]

Observed		Predicted		
		Persistence		Percent age Correct
		0 = not persisting	1 = persisting	
Persistence	0 = not persisting	157	37	80.9
	1 = persisting	99	140	58.6
Overall Percentage				68.6

a. The cut value is .500

While on the surface 68.6% may seem impressive, the classification table warrants a closer inspection. What is missing is information about the probability of the case classifications. Before the model can be deemed useful, a comparison of the accuracy rates must be undertaken.

Proportional by Chance Accuracy Criteria

The information in the classification table can be used to evaluate the utility of binary LR models by comparing the overall percentage correct with the proportion by chance accuracy criteria (PCC). This is computed by squaring and summing the proportion of cases for each group (Bayaga, 2010; El-Haib, 2012). To illustrate, consider the information in Tables 2-3. Upon initial inspection of two different student persistence models, White, Altschuld, and Lee (2006) and Mitchell (2011) found overall 74.6% and 73.8% correct classifications respectively. However, when proportion by chance was computed, both models failed to satisfy the criteria -- overall case classifications 25% higher than the proportion by chance rate. Thus the variables in the models examined by White and colleagues ($0.254^2 + 0.746^2 = 0.621 \times 1.25 = 77.6$) and Mitchell ($0.280^2 + 0.720^2 = 0.597 \times 1.25 = 74.6$) were not useful in predicting student persistence. Stated otherwise, the performance of the variables in the model was no better than could be reasonably expected by chance.

Table 2: Model classification table (*n = 311*)*

Classification Table[a]

Observed		Predicted		
		Persistence		Percentage Correct
		0 = not persisting	1 = persisting	
Persistence	0 = not persisting	11	68	13.9
	1 = persisting	11	221	95.3
Overall Percentage				74.6

a. The cut value is .500

* SPSS (Block 1: Method = Enter)

Table 3: Model classification table (*n = 1301*)*

Classification Table[a]

Observed		Predicted		
		Persistence		Percentage Correct
		0 = not persisting	1 = persisting	
Persistence	0 = not persisting	65	299	17.9
	1 = persisting	42	895	95.5
Overall Percentage				73.8

a. The cut value is .500

* SPSS (Block 1: Method = Enter)

What is missing from computation of proportion by chance accuracy is an examination of the case classifications before and after the predictor variables were entered into the regression equation. This calls for a comparison of the a priori and post priori classification tables to determine if the null model (constant) performed better. In the Table 1 example, 68.6% may seem impressive but most investigators are more interested in the accuracy of the predictions rather than goodness-of-fit.

Proportional Reduction in Error

There is no consensus on how to measure the association between the observed and predicted classification of cases in logistic regression. Menard (2002) recommends using the information from the classification tables to calculate the proportional change in error with a variant of the proportional reduction in error (PRE) statistic (Menard, 2004). The general principle is that knowing the value of the observed classification can be used to predict the value of the predicted using the formula E1 – E2/E1 where E1 = errors before the model and E2 = errors after the model. In contrast to the other aspects of logistic regression such as the Wald test of significance, chi square, and statistics analogous to R^2 where sample size is critical (Alam, Rao, & Cheng, 2010), it is not as important when analyzing classification tables. That is because the *n* value is not an element of the PRE formula.

In binary LR, all cases are predicted to belong to one of two possible outcomes: the event “occurring” or “not occurring”. When applied to the information in the classification tables, the PRE indicates the percent of fewer classification errors that will occur by using the variables in the logistic regression equation. In other words, this is a measure of the predictive accuracy of the model (Menard, 2004). Using information from the null and model classification tables, the proportional reduction in error is calculated as: E without the model – E with the model/E errors without the model. The PRE will vary between 0 and 1, indicating the efficiency of the model in predicting the occurrence or non-occurrence of the event. When the number of errors without the model equals the number with the model, the value will be 0. As an example, consider the *without the model* information in the classification table presented in Table 2 compared to the *with the model* data in Table 4. In examining student persistence White, Altschuld, and Lee (2006) found the same number of before (E1 = 79) and after errors (E2 = 79) even though they had an overall correct classification of 74.6%. In other words, the variables in the regression equation offered no additional predictive capability. In contrast, after reviewing the *without the model* classification data in Table 5, Mitchell (2011) found that his model of student persistence had more before (E1 = 364) than after errors (E2 = 341). This translated into a predictive efficiency of approximately 6.3%. However if the 73.8% overall correct classifications in Tables 3 are not scrutinized more closely, a different impression emerges of the model’s predictive ability.

Table 4: Null without the model classification table (*n = 311*)*

Classification Table[a]

Observed		Predicted		
		Persistence		Percentage Correct
		0 = not persisting	1 = persisting	
Persistence	0 = not persisting	0	79	0.0
	1 = persisting	0	232	100.0
Overall Percentage				74.6

a. The cut value is .500

* SPSS (Block 0: Beginning Block)

Table 5: Null without the model classification table (*n = 1301*)*

Classification Table[a]

Observed		Predicted		
		Persistence		Percentage Correct
		0 = not persisting	1 = persisting	
Persistence	0 = not persisting	0	364	0.0
	1 = persisting	0	937	100.0
Overall Percentage				72.0

a. The cut value is .500

* SPSS (Block 0: Beginning Block)

Closing Thoughts

Both the PCC and PRE techniques highlight the importance of going beyond the percentage of correct classifications to include a more thorough analysis. This paper demonstrates how the proportional by chance accuracy rate and proportional reduction in error statistic can be used to evaluate the effectiveness of binary logistic regression models (Long, 1997).

Finally, it illustrates the need for educational researchers not to become overly reliant on software. An explanation for this tendency may be the emphasis on methods that many cursory statistics courses have adopted in graduate education programs (Curran-Everett, Taylor, & Kafadar, 1998). None-the-less, what is critical is that educational researchers recognize that a fundamental knowledge of statistical concepts and principles, such as the ones discussed in this paper, is the cornerstone of scientific inquiry.

References

Alam, M.K., Rao, M.B., & Cheng, F.C. (2010). Sample size determination in logistic regression. *Sankhya: The Indian Journal of Statistics, 72*(1), 58-75.

Baradwaj, B.K. & Pal, S. (2011). Mining educational data to analyze students' performance. *Journal of Advanced Computer Science and Applications, 2*(6), 63-69.

Bayaga, A. (2010). Multinomial logistic regression: Usage and application in risk analysis. *Journal of Applied Quantitative Methods*, *5*(2), 288-297.

Cohen, J. (1983). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawarence Erlbaum Associates

Cox, D.R. & Snell, E.J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman & Hall.

Curran-Everett, D., Taylor, S., & Kafadar, K. (1998). Fundamental concepts in statistics: Elucidation and illustration. *Journal of Applied Physiology, 85*(3), 775-786.

El-Habil, A.M. (2012). An application on multinomial logistic regression. *Pakistan Journal of Statistics and Operation Research, 8*(2), 271-291.

Freese, J. & Long, J.S. (2006). *Regression models for categorical dependent variables using STATA.* College Station, TX: STATA Press.

Garson, G.D. (2004). Quantitative research in public administration. Retrieved May 30, 2013 from http://www2.chass.ncsu.edu/garson/pa765/logistic.htm

George, D. & Mallery, P. (2012). *IBM SPSS Statistics 19 Step by Step: A simple guide and reference (12th ed.).* Upper Saddle River, NJ: Pearson Education.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L. & Black, W.C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

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

Mitchell, S.K. (2011). *Factors that contribute to persistence and retention of underrepresented minority undergraduate students in science, technology, engineering, and mathematics-STEM* (Doctoral dissertation). University of Southern Mississippi, Hattiesburg, MS.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24.

Menard, S. (2002). *Applied logistic regression analysis.* Sage University Series on Quantitative Applications in Social Sciences, 07-106. Thousand Oaks, CA: Sage Publications.

Menard, S. (2004). Proportional reduction of error (PRE). In M. Lewis-Beck, A. Bryman, & T. Liao (Eds.), *Encyclopedia of social science research methods.* (pp. 877-878). Thousand Oaks, CA: SAGE Publications.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Norusis, M.J. (2002). *SPSS 12.0 statistical procedures companion.* Upper Saddle River, NJ: Prentice-Hall.

Pampel, F.C. (2000). *Logistic regression: A primer*. Sage University Series on Quantitative Applications in Social Sciences, 07-132. Thousand Oaks, CA: Sage Publications.

Peng, C.Y.J., Lee, K.L., & Ingersoll, G.M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14.

Stevens, J. (2007). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Teh, S.Y., Othman, A.R., & Michael, B.C. (2010). Dichotomous logistic regression with leave-one-out validation. *Proceedings of the World Academy of Science, Engineering and Technology, 62,* 538-547.

White, J.L., Altschuld, J.W. & Lee, Y.F. (2006). Persistence of interest in science, technology, engineering, and mathematics: A minority retention study. *Journal of Women and Minorities in Science and Engineering, 12*(1), 47-65.

Jeffry L. White is the Joan D. & Alexander S. Haig/BORSF Professor in Education at the University of Louisiana, Lafayette. He is co-author of *Needs Assessment: Analysis and Prioritization* (Sage, 2009) and his research interests are in quantitative methods and evaluation of retention programs for underrepresented groups in science, technology, engineering, and mathematics (STEM). **Dr. White** can be contacted at jwhite1@louisiana.edu