Accessing a Model's Ability to Classify Subjects: The Importance of Considering Marginally Accurate Classifications

Russell Brown
Cleveland State University

Isadore Newman
University of Akron

John W. Fraas
Ashland University

The purpose of this paper is fourfold: (a) discuss the importance of considering "marginally accurate" classifications, which are predicted probability values whose confidence limits contain the cut-value used to classifying subjects, (b) present a six-step calculation procedure used to identify the "marginally accurate" classification values, (c) illustrate how the identification of these "marginally accurate" values are important in the evaluation of the model, and (d) discuss how a review of the "marginally accurate" values can used to access the differential effectiveness of various modeling procedures with respect to their replicability and stability.

Fraas and Drushal (2004) suggested that program evaluators and educational researchers frequently encounter situations in which the dependent variable of interest is dichotomous (i.e., a variable that consists of two categories). One goal of analyzing a dichotomous dependent variable is to obtain a model that can be used to classify subjects into either of the two categories of the dependent variable. It is not uncommon for researchers and program evaluators to use a logistic regression model for this purpose. Fraas and Newman (2003) noted such classifications can be obtained from a linear probability regression model as well as a logistic regression model. In addition, Brown, Newman, and Fraas (2004), explain how a third-degree polynomial regression could be used to classify each subject into one of the two categories of the dependent variable.

Regardless of which analytic method is used, program evaluators and researchers need to consider an issue that is often overlooked. That is, how confident are they in the classification of the subjects? The purpose of this paper is fourfold: (a) discuss the importance of considering "marginally accurate" classifications, which are predicted probability values whose confidence limits contain the cut-value used to classifying subjects, (b) present a six-step calculation procedure used to identify the "marginally accurate" classification values, (c) illustrate how the identification of these "marginally accurate" values are important in the evaluation of the model, and (d) discuss how a review of the "marginally accurate" values can used to access the differential effectiveness of various modeling procedures with respect to their replicability and stability.

Need for Additional Classification Table Information

The need for providing information that can be used to supplement the classification table produced by analytic methods used in conjunction with a dichotomous dependent variable occurred to us when we attempted to compare the results of the three analytic techniques (Newman, Brown, & Fraas, 2004). We found that misclassifications became problematic in trying to explain the comparative results of three methods. Although similar results were obtained by the three different models when comparing the methods in terms of tests of significance and predicted probabilities, some differences existed in the group classifications produced by them. Under a condition in which there was a modest correlation between the independent variable, the third-degree polynomial model produced 3.5% more errors than did the logistic or linear models, which produced identical classification patterns. A closer examination of these differences in classification showed the cases that were classified differently had predicted probabilities that were all close to the cut-line probability level of .50.

We believe the degree of confidence we have in two models that assign the same classification to each subject may or may not be the same. If the first model produces predicted probabilities for the subjects that have greater variation than those of a second model and a number of those probabilities are located near the cut-line of .50, our confidence in the first model will not be as strong as it is in the second model even though both models classified the subjects the same.

To address this issue, we believe information conveyed by the classification table, which lists the number of subjects correctly classified and incorrectly classified, should be supplemented by reporting the number and percentage of classifications that are "marginally accurate." If the number or percentage of such classifications for a model is small, program evaluators and researchers would have greater confidence in using the model to classify future subjects. The discussion presented in the next section provides the steps researchers need to complete in order to access a model in such a fashion.

Method

A method is presented through an illustration that can be used to identify the number and percent of "marginally accurate" values (i.e., predicted student probability values whose confidence limits contain the cut-value used to classifying subjects). The illustration used is taken from Fraas and Drushal (2004) in their discussion of the use of delta-p values as a means to understand the effect of incremental changes in the independent variables on the predicted probability in a logistic model.

The data from the Fraas and Drushal (2004) study contained information on 525 college students. They were "interested in assessing the relationship between various student and financial factors recorded for students who have applied to a university and whether the students actually did or did not matriculate" (p. 5). The dependent variable indicated in which of two categories each student belonged. Each student who did not matriculate was assigned a value of one, while each student who did matriculate was assigned a value of zero. The independent variables used to predict whether or not an individual student did or did not matriculate were as follows:

1. The students' high school grade point averages (HSGPA)
2. The students' ACT composite scores (ACT)
3. The sex of each student (SEX) [0 = female student; 1 = male student]
4. The amount of financial aid offered each student (AID)
5. The amount of financial need established for each student (NEED).

A linear probability model, a third-degree polynomial model, and a logistic regression model were used to analyze the relationships between the independent variables and the dependent variable. In order to be able to produce the terms for the third-degree polynomial model, a multiple linear regression was used to produce a single standardized weighted predicted composite score for each subject. This score was then squared and cubed to produce the terms for the third-degree polynomial model.

Each regression method was subsequently used to establish predicted probabilities and predicted classifications for each of the subjects in terms of the dependent variable (i.e., was the subject predicted to matriculate or not matriculate). Subjects whose predicted probability values were greater than or equal to .5 were classified as having matriculated; while subjects whose predicted probability values were less than .5 were classified as not having matriculated. The classification results of these analyses can be seen in Table 1.

As one can see from the results listed in Table 1, the percent correctly classified by the three methods are quite similar. The polynomial model produced the greatest number of correct classifications (58.1%). The linear and logistic models produced an equal percentage of correct classifications (57.3%), but produced a different pattern of false positive and false negative identifications.

The specific issue we are attempting to address is: What number and percent of the classifications are "marginally accurate" (i.e., "unstable")? The calculation of the number and percent of correctly classified subjects whose classifications are "marginally accurate" can be calculated in six steps regardless of which model is used.

The calculations used in our illustration are for the logistic regression model. The required steps are as follows:

1. The two standard deviation values for the predicted probability values--one for the group of students who were classified by the model as not matriculating and the other for the group of students who were classified by the model as matriculating--are calculated. The standard deviation values for the groups of students who did not and did matriculate were .053 and .048, respectively.

2. Since we are interested in a one-tailed limit value for each group, the standard deviation value for each group is multiplied by 1.65 (the t value for the one-tailed 95% confidence level). Thus the value for the students who were classified by the model as not matriculating was .088 (.053 X 1.65), while the value for the students who were classified by the model as matriculating was .079 (.048 X 1.65).

Table 1. Original Group Membership Classifications and Errors

Model	Correct Classification		False Positives	False Negatives	Percent Correct
	1	0			
Linear Model	194	107	145	79	57.3%
Polynomial Model	185	120	132	88	58.1%
Logistic Model	195	106	146	78	57.3%

3. The value of .088 was added to each predicted probability for the students classified by the model as not matriculating; while .079 was subtracted to each predicted probability for the students classified by the model as matriculating.

4. The number of students who were classified by the model as not matriculating but whose upper predicted probability limit values equaled or exceeded .50 was recorded. A total of 35 students had upper limits that equaled or exceeded .50. Thus of the original 106 who were correctly classified as not matriculating, 33.0% had upper predicted probability limits that equaled or exceeded .50. We labeled these classifications as "marginally accurate."

5. The number of students who were classified by the model as matriculating but whose lower predicted probability limit values fell below .50 was recorded. A total of 71 students had lower limits that fell below .50. Thus of the original 195 who were correctly classified as not matriculating, 36.4% had lower predicted probability limits that fell below .50. Again, we labeled these classifications as "marginally accurate."

6. The total number and total percent of correctly classified students who were labeled as "marginally accurate" were noted. The total number was 106 (35 + 71) and the total percent was 35.2% [((35 + 71)/ 301) X 100].

The number and percent of students labeled "marginally accurate" or "unstable" were calculated in the same manner for the linear probability and the third-degree polynomial models. See Table 2 for the results of those calculations.

Regardless of which model is used, we suggest that the percent of "marginally accurate" figures (e.g., 33.0% of the students correctly classified as not matriculating; 36.4% of the students correctly classified as matriculating; and 35.2% of the students overall correctly classified for the logistic regression model) should be reported along with the classification table that is normal provided by an analysis of a dichotomized dependent variable.

An examination of the number of subjects "marginally accurate" for each of the three types of models may lead researchers to reach a different conclusion regarding the desirability of using a given model than would a review of the number of subjects correctly classified by each model is reviewed. The number of subjects correctly classified by each model (see Table 1) would suggest that the models are approximately equally effect in classifying students. A review of the number of subjects identified as "marginally accurate" would indicate that the polynomial model, which had the lowest number of "marginally classified" subjects (see Table 2) may be the preferred model. Thus it may be important, both in a relative and an absolute sense, for researchers to access both criteria (i.e., the number and percent "marginally accurate" as well as the number and percent correctly classified) when accessing a models ability to classify students.

Discussion

When attempting to evaluate the effectiveness of a model designed to classify subjects into one of two groups with a linear probability model, a third-degree polynomial model, or a logistic regression model researchers may find the information provided by the classification table insufficient. The application of the technique for identifying the number and percent of "marginally accurate" classifications, as presented in this paper, may be used to supplement the information presented in the standard classification table. The fewer the number of identified "marginally accurate" classifications the more confident the researchers will be in their model's ability to classify future subjects.

Table 2. Changes in Group Membership Classifications for Students Correctly Classified by the Model

Model and Original Correct Classification	Marginally Accurate	Stable	Total	% Marginally Accurate
Linear Model				
Matriculated	75	119	194	38.7%
Did not Matriculate	44	63	107	41.1%
Total	119	182	301	39.5%
Polynomial Model				
Matriculated	66	119	185	35.6%
Did not Matriculate	23	97	120	19.2%
Total	89	216	305	29.2%
Logistic Model				
Matriculated	71	125	195	36.4%
Did not Matriculate	35	70	106	33.0%

We believe that researchers often judge a model's ability to predict group membership or an occurrence of an event primarily through the use of the classification table. The values reported in the standard classification table, however, do not take into consideration the number of probability values (i.e., the values on which the classifications are based) that are close to the cut-value used to classify the subjects. If a large number of these probability values are located near the cut-value, researchers may find the accuracy of the classifications of future subjects unacceptable. That is, the model did not provide sufficient stability from sample to sample.

Researchers may find it important to identify in which classification most of the "marginally accurate" values are located (i.e., in the classification assigned the value of 0 versus the classification assigned the value of 1). If the marginal values are predominately in the classification assigned the value of 1 (the event did occur) and few or no marginal values are located in the classification assigned the value of 0 (the event did not occur), the researchers may be more confident in their classifications of the event not occurring than not occurring.

We realize there are other methods that researchers can use to access a model's ability to classify subjects (e.g., the use of a holdout group). The key issue, however, is that regardless of how researchers evaluate a model's ability to classify subjects, consideration should be given to the confidence they have in their classifications. We believe is an analytic technique that will improve data-based decision making by forcing researchers to reflect on how their models will be used and the degree of confidence they have in the use of those models. If a model is to be used to classify future subjects, the concept of "marginally accurate" values may be a key concept for researchers to consider.

References

Brown, R., & Newman, I. (2002). A discussion of an alternative method for modeling cyclical phenomena. *Multiple Linear Regression Viewpoints, 28*(1), 31-35.

Fraas, J. W. & Newman, I. (February, 2003). *Ordinary lest squares regression, discriminant analysis, and logistic regression: Questions researchers and practitioners should address when selecting an analytic technique*. Paper presented at the Eastern Educational Research Association, Hilton Head, SC.

Fraas, J. W., & Drushal, J. (2004). *Expressing logistic regression coefficients as delta-p values.* Manuscript presented for publication.

Newman, I., Brown, R., & Fraas, J. W. (April, 2004). *Logistic regression as compared to linear and polynomial least squares regression: Is OLS 3rd degree polynomial a more fair comparison to the logistic method.* Paper presented at the American Educational Research Association, San Diego, CA.

Send correspondence to: Russell Brown, Ph.D.
XXXXXX
Cleveland State University
Cleveland, OH 76203-1335
Email: xxxxxxxxx