Research in Higher Education
Journal of the Association for Institutional Research

Volume 43, Number 3 June 2002

The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999

Chao-Ying Joanne Peng, Tak-Shing Harry So, Frances K. Stage, and Edward P. St. John ........................................ 259

Examining the Institutional Transformation Process: The Importance of Sensemaking, Interrelated Strategies, and Balance

Adrianna Kezar and Peter Eckel ........................................ 295

Curriculum Leadership Roles of Chairpersons in Continuously Planning Departments

Joan S. Stark, Charlotte L. Briggs, and Jean Rowland-Poplawski ........................................ 329

Honor Codes and Other Contextual Influences on Academic Integrity: A Replication and Extension to Modified Honor Code Settings

Donald L. McCabe, Linda Klebe Treviño, and Kenneth D. Butterfield ........................................ 357

RESEARCH AND PRACTICE—Measuring the Quality of Faculty and Administrative Worklife: Implications for College and University Campuses

Linda K. Johnsrud ........................................ 379

Chao-Ying Joanne Peng, Tak-Shing Harry So, Frances K. Stage, and Edward P. St. John

This article examines the use and interpretation of logistic regression in three leading higher education research journals from 1988 to 1999. The journals were selected because of their emphasis on research, relevance to higher education issues, broad coverage of research topics, and reputable editorial policies. The term "logistic regression" encompasses logit modeling, probit modeling, and tobit modeling and the significance tests of their estimates. A total of 52 articles were identified as using logistic regression. Our review uncovered an increasingly sophisticated use of logistic regression for a wide range of topics. At the same time, there continues to be confusion over terminology. The sample sizes used did not always achieve a desired level of stability in the parameters estimated. Discussion of results in terms of delta-Ps and marginal probabilities was not always cautious, according to definitions. The review is concluded with recommendations for journal editors and researchers in formulating appropriate editorial policies and practice for applying the versatile logistic regression technique and in communicating its results with readers of higher education research.

KEY WORDS: logistic regression; logit; probit; tobit; delta-P; marginal probability; odds ratio; higher education research; multivariate statistics.

Since 1988, research using logistic regression has been published with increasing frequency in three leading higher education journals: Research in Higher Education, The Review of Higher Education, and The Journal of Higher Education. Yet, there is great variation in the presentation and interpretation of results in these publications, which can make it difficult for readers to understand and compare the results across articles. A systematic review of articles that have used logistic regression not only promotes the learning about this method, but


Address all correspondence to: Chao-Ying Joanne Peng, Department of Counseling and Educational Psychology, School of Education, Room 4050, 201 N. Rose Ave., Indiana University, Bloomington, IN 47405-1006; peng@indiana.edu.

259
also helps suggest new guidelines for principled applications of this versatile technique.

Logistic regression, being one special class of regression models, is well suited for the study of categorical outcome variables, such as staying in or dropping out from college. This technique is increasingly applied in educational research. A search of the ERIC database indicated that between January 1988 and December 1999, a keyword search on terms such as “logistic regression,” “logit,” “probit,” “normit,” or “tobit” produced 90 abstracts that investigated education issues, out of a total of 233 (or 38.63%). The proportion of higher education-related articles increased noticeably beginning in 1992. In recent years, the application of logistic regression has been found more frequently in the annual meeting programs of the Postsecondary Education Division of the American Educational Researcher Association than in those of the other 11 divisions. The trend in higher education is for researchers to recognize limitations of ordinary least squares (OLS) regression and turn increasingly to logistic regression for explaining relationships between a categorical outcome variable and a mixture of continuous and categorical predictors. This trend is primarily motivated by complex data and categorical outcome measures, for example, enrollment/matri
culation, retention, and graduation, that are of interest to higher education researchers. It is also facilitated by four excellent papers written by Hinkle, Austin, and McLaughlin (1989), Austin, Yaffee, and Hinkle (1992), Dey and Astin (1993), and Cabrera (1994) on logistic regression and its variations such as, log-linear models, probit models, and linear probability models.

Despite the popularity of logistic regression in recent years, confusion continues to exist over terms, concepts, practices, and interpretations. Logistic regression results have been reported in terms of logit, odds, odds ratio, relative risk, predicted probability, marginal probability (also called marginal effect, partial effect, or partial change), and change in predicted probability (also called delta-
\( P \)). These terms are not equivalent; thus, their use is not interchangeable. Which of these eight is the most suitable to report? The answer depends on the nature of data and purposes of the study. Likewise, a researcher can employ at least four modeling techniques to fit a logistic regression model to the data: direct, sequential, stepwise, and best \( k \)-predictors modeling. Further, the fit between each model and the data may be assessed by various indices; some are more appropriate than others for certain types of models. This, too, causes confusion among researchers.

While the research community recognizes the superiority of logistic regression over OLS models, two questions still remain: “Are all logistic regression analyses conducted appropriately and comprehensively?” and “Are the results reported informatively that allow for comparisons across studies?” To help answer these two questions, we decided to systematically review the use of logistic regression in higher education research. Specifically, we focused on articles
published between 1988 and 1999 in three leading higher education research journals: *Research in Higher Education*, *The Review of Higher Education*, and *The Journal of Higher Education*. These three journals were cited in Silverman (1985) and Budd (1988) as core journals in higher education that publish research on a broad range of issues in higher education (Budd, 1988, p. 181). It is hoped that such a review informs editors and researchers in formulating and promoting standards for the application of logistic regression. Many important student outcomes in higher education—including first-time enrollment and persistence decisions—are appropriately conceptualized as dichotomous outcomes. Students either choose to enroll, or continue their enrollment, or do not. However, since graduate students in higher education are more likely to take graduate courses that cover OLS regression than logistic regression, it is imperative that the complexities of logistic modeling be well understood within the higher education research community.

The remainder of this article is divided into four sections: (a) Logistic Regression Models, (b) Review Objectives and Method, (c) Results and Discussion, and (d) Summary and Recommendations. The appendix lists all articles we reviewed.

**LOGISTIC REGRESSION MODELS**

In this section, we first review basic concepts in logistic regression including the logistic function, the polytomous model, the logit model, the probit model, and the tobit model. This is followed by discussions of various analysis and interpretation issues for which standards are established for the subsequent review of the application of logistic regression. Specifically, we discuss the minimum observation to predictor ratio, examination of possible interaction effects, evaluations of logistic regression models, diagnostic statistics, and three reporting formats of logistic regression results.

A typical regression model has the following general appearance

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \]  \hspace{1cm} (1)

where \( \hat{y} \) is the estimated outcome variable value for the true \( Y \), \( b_0 \) is the constant of the equation, \( b_1, \ldots, b_p \) are estimated parameters corresponding to predictor values \( x_1, \ldots, x_p \); \( b_0 \) is alternatively called the \( Y \)-intercept; \( b_1, \ldots, b_p \) are slopes, regression coefficients, or regression weights.

One method used by statisticians to estimate parameters is the least squares method. The values obtained under the least squares method are called least squares estimates. In the case of categorical outcome variables (such as whether a high school graduate matriculated into college), the linear regression model is inadequate. For one thing, the plot of such data appears to fall on parallel lines,
each corresponding to a value of the outcome variable, say, 1 = matriculated and 0 = did not matriculate. Hence, the variance of residuals at specific values of \( X \) (say, GPA = 3.50 or GPA = 2.5) equals \( p \times (1 - p) \), where \( p \) is the proportion of a particular outcome (such as matriculated) at specific \( X \) values. The variance is obviously a function of \( p \) rather than a constant. Furthermore, the categorical nature of the outcome makes it impossible to satisfy either the normality assumption for residuals or the continuous, unbounded assumptions on \( Y \). As a result, significance tests performed on regression coefficients are not valid, although least squares estimates are unbiased (Menard, 1995). Even if the categorical outcomes are reconceptualized as probabilities, the predicted probabilities derived from least squares regression models can sometimes exceed the logical range of 0 to 1. This results from a lack of provision in the model to restrict the predicted values. Further, the \( R^2 \) index derived from least squares regression for categorical outcomes does not render the usual meaning of variance explained (Menard, 2000); it does not correspond to the predictive efficiency and cannot be tested in an inferential framework (Menard, 2000).

To overcome the limitations of least squares regression in handling categorical variables, a number of alternative statistical techniques have been proposed. These include: logistic regression, discriminant function analysis, log-linear models, and linear probability models. Compared to the other three alternative techniques, logistic regression is superior because it (a) can accept both continuous and discrete predictors, (b) is not constrained by normality or equal variance/covariance assumptions for the residuals, and (c) is related to the discriminant function analysis through the Bayes theorem (Flury, 1997, p. 558). Furthermore, in terms of classification and prediction, logistic regression has been shown to produce fairly accurate results (Fan and Wang, 1999; Lei and Koehly, 2000). For these reasons, researchers in higher education have recognized logistic regression as a viable alternative to linear discriminant function analysis and other techniques for analyzing categorical outcome variables.

In the simplest case of one predictor \( X \) and one dichotomous outcome variable \( Y \), the logistic regression model predicts the logit of \( Y \) from \( X \). The logit is the natural logarithm (ln) of odds of \( Y = 1 \) (the outcome of interest). The simple logistic model has the form:

\[
\ln \left( \frac{P}{1 - P} \right) = \log(\text{odds}) = \logit = \alpha + \beta x \tag{2}
\]

Hence,

\[
\text{Probability} \ (Y = \text{outcome of interest} \mid X = x) = P = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = g(x). \tag{3}
\]
where $P$ is the probability of the outcome of interest (or the “event”) under variable $Y$, $\alpha$ is the $Y$ intercept, and $\beta$ is the slope parameter. Both the $Y$ intercept and the slope parameter are estimated by the maximum likelihood (ML) method. The ML method is designed to maximize the likelihood of obtaining the data given its parameter estimates. As Equation 3 illustrates, the relationship is non-linear between parameters ($\alpha$, $\beta$) and the probability of observing a particular outcome in an observation (such as a high school graduate matriculating). Yet the relationship between ($\alpha$, $\beta$) and the logit is linear. For a single predictor $X$, the logistic regression curve looks like a normal ogive (Figure 1 for $\alpha = 0$ and $\beta = 0.16428$).

Within the inferential framework, the null hypothesis states that $\beta$ equals zero in the population. Rejecting such a null hypothesis implies that a relationship exists between $X$ and $Y$. If the predictor is binary, such as gender, the exponentiated $\beta$ ($= e^\beta$) is simply the odds ratio, or the ratio of two odds. In the example of relating students’ matriculation to their gender, one can interpret such a relationship using the concept of odds ratio. Say, the odds for girls to matriculate is 1.5 times more (or less) likely than the odds for boys. Peterson (1985) suggests that delta-$P$ (or change in the probability) be used in interpreting the logistic regression result if the predictor is continuous. These two interpretations are straightforward and have been employed by authors of the 52 articles.

**FIG. 1.** Univariate logistic regression model based on $\alpha = 0$ and $\beta = 0.16428$. 
The logistic function, that is, the \( g(x) \) in Equation 3, has the following unique characteristics:

1. Unless \( \beta = 0 \), the binary logistic regression maps the regression line onto the interval \((0,1)\) which is compatible with the logical range of probabilities.
2. The regression line is monotonically increasing if \( \beta > 0 \), and monotonically decreasing if \( \beta < 0 \).
3. The function takes the value of 0.5 at \( x = -\alpha/\beta \) and is symmetric to the point of \((-\alpha/\beta, 0.5)\).

As these properties demonstrate, the logistic regression model guarantees that (a) the predicted probabilities \( \hat{P} \) will not fall outside the range of 0 to 1; (b) the slope parameter has the same meaning as the slope parameter in least squares regression models; and (c) the logistic function has a point of inflection corresponding exactly to 50% probability.

Using the same logic as that underlying the simple logistic regression, a complex model can be constructed to improve the prediction of the logit by including several predictors. The complex logistic model is in the same form as multiple regression equations. It is given by:

\[
\ln \left( \frac{P}{1 - P} \right) = \text{logit} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \tag{4}
\]

Therefore,

\[
\text{Probability}(Y = \text{outcome of interest} \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = P = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}. \tag{5}
\]

where \( P \) is once again the probability of the "event" under the outcome variable \( Y \), \( \alpha \) is the \( Y \) intercept parameter, \( \beta s \) are slope parameters, and \( Xs \) are a set of predictors. Once again, \( \alpha \) and \( \beta s \) are estimated by the ML method. The interpretation of \( \beta s \) is rendered using either the odds ratio (for categorical predictors) or the delta-\( P \) (for continuous predictors). The null hypothesis states that all \( \beta s \) equal zero. A rejection of this null hypothesis implies that at least one \( \beta \) does not equal zero in the population.

If the dependent variable \( Y \) can take on any of several possible outcomes that are ordered, then polytomous logistic regression is called for. Suppose, for example, three outcomes are possible with \( Y \): 1 = matriculating in college, 2 = matriciating on probation, and 3 = involuntary withdrawal. Define
\[ P_1 = \text{Prob} \left( Y = 1 \mid X \right), \]
\[ P_2 = \text{Prob} \left( Y = 2 \mid X \right), \text{ and} \]
\[ P_3 = \text{Prob} \left( Y = 3 \mid X \right), \text{ where } X = \text{vector of predictors } X_1, X_2, \ldots, X_k. \]

Then the polytomous logistic regression models \( P_1 \) and the cumulative probability of \( P_1 + P_2 \). Thus, both \( P_2 \) and \( P_3 \) are derived by subtraction. The parameters may be allowed to be different or constrained to be identical. If they are constrained to be equal, then the resulting logistic regression models yield common slopes. This type of polytomous model is referred to as the proportional-odds model since the ratio of the odds of the event \( Y \leq j \) is independent of the category, \( j \). In other words, the odds ratio is constant for all categories.

The logistic regression model is also referred to as the logit model due to the logit transformation of \( P \), that is, \( \ln \left( \frac{P}{1 - P} \right) \), that leads to a logit scale with a mean of 0 and variance of \( \pi^2/3 \). At least one other function (probit), besides the logit, may be used by researchers to link the predicted probability of the “event” outcome to a set of predictors (Dey and Astin, 1993; SAS Institute Inc., 1999). The probit function is also called normit, which has the following form:

\[ \text{Probit} = \Phi^{-1}(P) = \text{standard normal } z \text{ score} \]
\[ = \alpha + \beta x \text{ (in the simple probit model).} \tag{6} \]

The probit function is simply the inverse of the cumulative standard normal distribution function that yields the standard normal \( z \) score. In other words,

\[ \Phi(x) = \sqrt{2\pi}^{-1} \int_{-\infty}^{x} \exp(-z^2/2)dz = P. \tag{7} \]

The \( \Phi(x) \) function is comparable to the \( g(x) \) in Equation 3 for the logit. The probit (or normit) scale understandably has a mean of 0 and a variance of 1.

The choice of a particular link function depends on the underlying relationship between the outcome variable and the predictors: cumulative normal or logistic. When comparing regression coefficients obtained from different functions, researchers need to be aware of the scaling differences due to the link function. In general, results obtained from logit and probit models are similar. They differ by approximately \( \sqrt{\pi^2/3} = 1.8138 \), with the logit coefficients larger than the probit coefficients by this ratio. This is especially true if the predicted probabilities are in the range of 0.1 to 0.9. Thus, both models may be used interchangeably.

The tobit model has the same mathematical model as the probit; the probabil-
ity of an outcome is derived from the cumulative normal distribution function. Hence, the inverse of the cumulative normal distribution function, the $z$ score, is explained by a set of predictors and their corresponding slopes. The major difference between tobit and probit models is in the nature of data to which each model is applicable. The tobit model applies to censored data, while probit applies to uncensored data. Censored data occur when a researcher has limited information about the outcome of certain cases in a sample. For example, a student registered zero credit hours for a coming semester in a retention study. Since the reason for registering zero credit hours could be due to the student’s (a) voluntary withdrawal, (b) involuntary withdrawal, or (c) other reasons, the outcome is not clear and cannot be determined exactly. We consider such a case as censored. Tobit is applied to data for which there are a substantial number of censored cases, the dependent variable cannot fall below a threshold (such as zero credit hour registered), and the residuals are normally and independently distributed.

**Minimum Observation to Predictor Ratio**

Similar to other statistical models, a logistic regression model derived from the sample is subject to sampling errors. Consequently, estimates for the regression coefficients become unstable for small samples. What is the recommended minimum for the observation/predictor ratio? This question cannot be answered directly since the literature has not offered specific rules applicable to logistic regression. Most rules we found were presented within the context of the ML method, for which the desirable properties of consistency, normality, and efficiency are asymptotic. Furthermore, in terms of statistical significance tests, the greater the sample size, the better the $\chi^2$ approximation to the sampling distribution of ML estimators. Since ML is the method of choice for estimating logistic regression coefficients, we decided to rely on these rules to assess the adequacy of sample sizes used in published research.

As Long (1997) pointed out, “it is risky to use ML with samples smaller than 100, while samples over 500 seem adequate” (p. 54). He further stated: “a rule of at least 10 observations per parameter seems reasonable for the models in this book” (p. 54). Lawley and Maxwell (1971) suggested that the significance test of the ML factor analysis solutions is appropriate if the sample contains at least 51 more cases than the number of variables under consideration. That is, $N - k - 1 \geq 50$, where $N$ is the sample size and $k$ is the number of variables (or predictors in the case of logistic regression). This is only a general rule of thumb, as Kim and Mueller (1978) correctly noted.

Within the context of canonical analyses and the ML method, Thorndike (1978) formulated two rules of thumb regarding sufficient sample size. His first rule was that sample size should be at least 50, plus 10 times the number of variables. The rule results in a sample size similar to the minimum ratio (10 to 1) recommended by Marascuilo and Levin (1983), except for small data sets.
Thorndike's second rule of thumb was that the sample size should be equal to 50, plus the squares of the number of variables. This second rule results in a required sample size that increases rapidly as the number of variables increases.

If data are ill-conditioned, for example, the predictors are highly linearly correlated or the baseline proportion of the outcome of interest is extreme, that is, $Y$ is nearly all 1's or all 0's, a larger sample is required. The ordered polytomous logistic model seems to require more observations (Long, 1997, p. 54). For stepwise logistic regression, a procedure notorious for capitalizing on chance, a larger ratio of observations to predictors should be expected. Tabachnick and Fidell (1996, p. 133) recommended a ratio of 40 to 1 for stepwise OLS multiple regression.

Although the minimum observation/predictor ratio to achieve stability of coefficients varies across authors cited here, several authors recommended a minimum ratio of 10 to 1 with a minimum sample size of 100 or 50 plus a variable number that is a function of the number of predictors. Researchers in general agreed that (a) the ML estimators of logistic regression coefficients would be stabilized with large samples, (b) certain regression models and/or data structures seem to require even larger samples, and (c) a conservative significance level should be adopted as evidence against the null hypothesis in small samples (Allison, 1995, p. 80; Long, 1997, p. 54).

Examination of Possible Interaction Effects

Examining the possibility of interaction between predictors is an essential step in model building strategies. As stated in Hosmer and Lemeshow (1989): "[Step] (4) Once we have obtained a model that we feel contains the essential variables, we should look more closely at the variables in the model and consider the need for including interaction terms among the variables" (p. 88).

The examination of interactions should be conducted after the appropriate scales of continuous predictors and the proper categories of discrete predictors have been ascertained. Hosmer and Lemeshow (1989) commented that

> [o]nce we have ascertained that each of the continuous variables is in the correct scale, we begin to check for interaction in the model. . . . The need to include interaction terms in a model is assessed by first creating the appropriate product of the variables in question and then using a likelihood ratio test to assess their significance. By significance, we mean interactions must contribute to the model. . . . In general, for an interaction term to alter both point and interval estimates, the estimated coefficient for the interaction term must attain at least a moderate level of statistical significance. The final decision as to whether an interaction terms should be included in a model should be based on statistical as well as practical considerations. (p. 91)

Evaluations of Logistic Regression Models

Once a logistic model is formulated, its adequacy is evaluated by a variety of statistical tests and indexes. These include: (a) tests of individual parameter
estimates in the model, (b) tests of the overall model, (c) validation of predicted probabilities, and (d) goodness-of-fit statistic. Individual parameter estimates are tested by the likelihood ratio test, the Wald statistic, or the Score test. According to Jennings (1986), Long (1997), and Tabachnick and Fidell (1996), the likelihood ratio test is more powerful than the Wald test, while the Score test is a normal approximation to the likelihood ratio test. A logistic model is said to provide an overall fit to the data if it demonstrates an improvement beyond the intercept-only model (also called the null model). Such an improvement is examined by three tests: the likelihood ratio, Score, and Wald tests.

Validation of predicted probabilities is typically presented in terms of percentages of correct classifications, Somers' $D$ statistic, sensitivity, specificity, false positive, false negative, or concordance pairs. The goodness-of-fit statistic is reported either as chi-square tests or the $R^2$ type of indexes. The chi-square test is based on Hosmer and Lemeshow statistic or the deviance. The Hosmer and Lemeshow statistic is both conservative and sensitive to the way in which predicted probabilities are grouped (Hosmer and Lemeshow, 1989, pp. 143–144; Ryan, 1997, p. 279). Deviance-based chi-square tests are correct only if they are based on the difference between two deviances rather than a deviance alone (McCullagh and Nelder, 1989). Further, deviance-based chi-square tests should be computed from covariate patterns, rather than from individual observations. For this reason, the deviance-based chi-square tests computed by SPSS and SYSTAT are incorrect (Peng and So, 1999, 2002).

The $R^2$ type of goodness-of-fit indexes takes on numerous forms; none should be interpreted as the percent of variance in the outcome variable that is explained by predictors (Long, 1997, pp. 104–109; Menard, 2000). Among the various $R^2$ analogs proposed for logistic regression, McFadden's index is preferred over others for its conceptual similarity to the OLS coefficient of determination, its relative independence from the base rate, and its comparability across models comprised of different predictors yet applied to the same outcome variable and the same data (Menard). It is defined as the difference between the initial and model $-2$ log-likelihood statistics, divided by the initial $-2$ log-likelihood statistic (McFadden, 1973). It is worth noting that the McFadden $R^2$ is not necessarily linearly related to the percentage of correct classifications in empirical studies (Menard).

Diagnostic Statistics

Diagnostic statistics in logistic regression reveal how influential each observation is to the fit of the logistic regression model. In 1981, Pregibon developed a series of diagnostic statistics that measure the influence of observations or identify outliers (i.e., ill-fitted observations). Technically speaking, these statistics are computed using covariate patterns. Covariate patterns are grouped data
for which the total number of observations having identical values on all predictors (also called covariates) is recorded along with the number of observations that yielded an "event" outcome on the dependent variable. Thus, diagnostic statistics are used to identify these patterns of observations that either exercise an exceedingly large influence over the fit of the model or whose predicted probabilities do not match their outcomes on the dependent variable. The examination of diagnostic statistics is meaningful and appropriate when the number of unique covariate patterns is much smaller than the number of observations (Hosmer and Lemeshow, 1989).

Seven diagnostic statistics may be obtained from statistical packages, such as SAS, SPSS, SYSTAT: (a) Pearson and (b) deviance residuals, (c) change in the Pearson chi-square statistic and (d) change in the deviance, (e) the change in parameter estimates due to a particular covariate pattern deleted, (f) the hat matrix diagonal, and (g) confidence interval displacement diagnostics.

Reporting Logistic Regression Results: Delta-\(P\), Marginal Probability, or Odds Ratio

Delta-\(P\) statistic, marginal probability, and odds ratio are commonly used to express logistic regression results. These three distinct concepts are not interchangeable. The delta-\(P\) statistic was popularized by an article written by Peterson (1985) in which the concept of delta-\(P\) was mathematically defined. According to Peterson, the delta-\(P\) means the change in probability as a result of a unit change in an independent predictor, say \(X_j\). Thus, delta-\(P\) (\(\Delta P\)) equals

\[
\Delta P = P(Y = 1 \mid L_1) - P(Y = 1 \mid L_0)
= P(Y = 1 \mid L_0 + \beta_j) - P(Y = 1 \mid L_0).
\]

(8)

The quantity \(L_1\) represents the logit after a 1-unit change in \(X_j\), whereas \(L_0\) represents the logit before the 1-unit change in \(X_j\). Both \(L_1\) and \(L_0\) are computed from Equation 4 as a result of a one-unit change in \(X_j\) while holding constants for all other predictors. It is evident from Equation 8 that delta-\(P\) is a function of both \(L_1\) and \(L_0\). In other words, the magnitude of delta-\(P\) is not a constant but rather a variable for the entire range of \(X_j\). If \(X_j\) is continuous, the delta-\(P\) changes as a function of both \(\beta_j\) (the regression coefficient of \(X_j\)) and \(L_0\). If \(X_j\) is categorical, the magnitude of delta-\(P\) depends on \(\beta_j\) and the reference category of \(X_j\). Equation 5 shows that the relationship is not linear between the probability of observing the outcome of interest and the predictors. Thus, it is incorrect to convert a regression coefficient, \(\beta_p\), to delta-\(P\) and then interpret the delta-\(P\) value without regard to the value of \(L_0\) (in the case of a continuous predictor) or the reference category (in the case of a categorical predictor).
Marginal probability (also called marginal effect, partial effect, or partial change) is another way of discussing logistic regression results. The technical definition of marginal probabilities for dichotomous outcomes is given in Long (1997) as: “the partial derivative of Equation [5 in this article] with respect to \( x_i \), [a specific value of predictor \( X_i \)]” (pp. 71–72). Thus, the marginal probability is the slope of the probability curve relating \( x_i \) to \( \Pr(Y = 1 \mid X_1 = x_1, \ldots, X_i = x_i) \), holding all other predictors constant. The sign of the marginal effect is determined by \( \beta_i \). Notice how the marginal probability is conditioned on the logistic regression model being realized on all predictors.

Since the magnitude of the marginal probability depends on the levels of all predictors, a researcher must decide which levels (or values) of the predictors to use when computing the marginal probability. One method is to compute the marginal probability at the mean of all predictors; another is to compute the average over all observations. The marginal probability at the mean is a popular summary measure for logistic regression models. It is, by Long’s observation (1997), “frequently included in tables presenting results, and is automatically computed by programs such as LIMDEP” (p. 74). One drawback, however, is that average values of predictors may not correspond to any observed values in the population.

The concept of marginal probability is not useful for explaining logistic regression models, regardless of which method is used. Two reasons are given in Long (1997, pp. 74–75). First, the measure is inappropriate for a binary or categorical predictor. Second, given the nonlinear relationship between probabilities and predictors, it is difficult to translate the marginal probability into the change in the predicted probability that will occur if there is a discrete change in \( X_i \). The difficulty is particularly pronounced if the range of a predictor corresponds to a region of the probability curve that is nonlinear. In this case, marginal probabilities should not be used as an indicator of predictors’ relative effects on the outcome.

A preferred alternative to the marginal probability for interpreting logistic regression results is delta-\( P \), when it is computed for a unit change in a predictor. These two concepts are not identical, as Figure 2 indicates. The marginal probability is the tangent at \( x_i \), and its value is the length of the vertical side of the solid triangle. For comparison purposes, let the change in \( x_i \) be one unit. The delta-\( P \) measures the change in the probability as \( x_i \) increases to \( x_i + 1 \). This is represented by the vertical side of the dash-lined triangle. Thus, the delta-\( P \) and the marginal probability are conceptually and numerically different. Only when the change in \( x_i \) occurs over a roughly linear region of the probability curve will the two measures be similar.

Odds ratios are directly derived from regression coefficients in logistic modeling. If \( \beta_j \) represents the regression coefficient for predictor \( X_j \), then exponentiating \( \beta_j \) yields the odds ratio. When all other predictors are held at a constant, the
odds ratio means the change in the odds of \( Y \), given a unit change in the \( X_i \). It is one of three epidemiological measures of effect that have been recently recommended by psychologists for informing public policymakers (Scott, Mason, and Chapman, 1999).

Three conditions must be met before odds ratios can be interpreted sensibly: (a) the predictor \( X_i \) does not interact with another predictor; (b) the predictor \( X_i \) is represented by a single term in the model; (c) a one-unit change in the predictor \( X_i \) is meaningful and relevant. It is worth noting that odds ratio and odds are two different concepts. They are related but not linearly related. Likewise, the relationship between the predicted probability and odds, although positive, is not linear either.

REVIEW OBJECTIVES AND METHOD

The application of logistic regression published in three leading higher education research journals—Research in Higher Education, The Review of Higher Education, and The Journal of Higher Education—was reviewed. The review period included 1988 to 1999. Within this period, a total of 52 articles were found to have used logistic regression. The criterion used in selecting articles was simple: at least one empirical analysis in the article must have been conducted to derive the logistic model and its regression coefficients. Any variation
of logistic regression—such as tobit, probit, ordered, or unordered polytomous logistic regression—qualified an article for our review. This criterion excluded any article that relied on others' work to derive the model or merely performed a logarithm or logit transformation of the dependent or the independent variable. A complete list of these 52 articles is found in the appendix. According to this list, 13 authors used logistic regression more than once.

Between 1988 and 1999, at least one article using logistic regression was published each year, most often two or more per year. Of all articles published in the three journals, 5.17% employed logistic regression. This percentage varied across the three journals from 8.49% (36 out of 424) for Research in Higher Education, to 2.82% (7 out of 248) for The Review of Higher Education, and 2.70% (9 out of 333) for The Journal of Higher Education.

The breakdown of articles by years showed that there has been an increasing use of logistic regression in higher education research, especially after 1992. This phenomenon may be explained by the advent of high-speed computers and the augmented availability of statistical software needed to perform the analysis. The majority of these articles appeared in Research in Higher Education (69.23%, or 36 out of 52), followed by The Journal of Higher Education (17.31%, or 9 out of 52), and The Review of Higher Education (13.46%, or 7 out of 52).

Our review was guided by three articles (Austin et al., 1992; Cabrera, 1994; Dey and Astin, 1993) that laid the theoretical groundwork for logistic regression. In addition, multivariate statistics books authored by DeMaris (1992), Hosmer and Lemeshow (1989), Kleinbaum (1994), Long (1997), McCullagh and Nelder (1989), and Menard (1995) were consulted during the review process. The review article written by Pugh and Hu (1991) on the use and interpretation of canonical correlation analyses and the coding form developed by Hossler and Scalese-Love (1989) for grounded meta-analysis of quantitative research were helpful in developing our own coding form. From these 11 sources and a related article on log-linear models by Hinkle et al., (1989), we formulated eight questions that were used in reviewing the 52 articles:

1. What was/were research questions of the study?
2. How many observations were used? How many predictor variables were included in each study? What was the observation/predictor ratio?
3. Was preliminary analysis conducted to investigate the need to transform measurement scales?
4. What modeling technique was used to derive the model? What indexes/coefficients were used in assessing the model? What statistical package was used in deriving the model and its related statistics?
5. Was any possible interaction or confounding examined?
6. Was diagnostic analysis of outliers performed? If so, what type of data was used?
7. How was the result of logistic regression reported and interpreted?
8. Was a multivariate statistics reference cited? If so, what was it?

Each article was reviewed independently by at least two authors. Differences in answers to the same question for a study were resolved through discussion.

RESULTS AND DISCUSSIONS

Among the 52 articles reviewed, only one by Dey and Astin (1993) performed both the logit and probit analyses. Tobit modeling was used by Bivin and Rooney (1999); it was mentioned in Lindahl and Winship (1994), though not used. The remaining 50 articles applied the logistic (or logit) model to data; of these, three implemented polytomous logistic modeling.

Guided by the eight questions posited earlier, we reviewed each article independently. Results for each question are as follows.

1. What Was/Were Research Questions of the Study?

The research questions addressed by logistic regression models fell into 13 types. More than half of the studies (29, or 55.77%) were related to university enrollment and retention. One explanation for this phenomenon is that, in these types of studies, the outcome measures considered were typically dichotomous or categorical, and at least one predictor was also categorical. Therefore, logistic regression was a suitable analytical tool. Another explanation is that authors of most of these studies were keen on applying the logistic regression technique for the research they pursued. Many of them applied logistic regression repeatedly, as alluded to earlier. Other types of research included: admission/application/attendance at college or graduate school (10 studies, or 19.23%), achievement/academic progress (5 studies, or 9.62%), university faculty, loan default (3 studies each, or 5.77%), charitable giving, student developments, diversity, institutional structure/effectiveness (2 studies each, or 3.85%), college rankings, admission statements and curricular, firm utilization of university scientific research, and medical career aspiration (1 study each, or 1.92%).

2. How Many Observations Were Used? How Many Predictors Were There in Each Study? What Was the Observation/Predictor Ratio?

In our review of 52 studies, we found the ratio of observations to predictors varied widely (Table 1) from 2728.8:1 (Lindahl & Winship, 1994) or 1773.3:1 by Strenta, Elliott, Adair, Matier, and Scott (1994) to 3.33:1 (Grubb, 1989, for women's data). For 9 of the 166 analyses (or 5.42%), the observation/predictor ratio was below 10:1—a minimum recommended by Long (1997, p. 54) for
<table>
<thead>
<tr>
<th>Articles</th>
<th>Number of Observations</th>
<th>Number of Predictors</th>
<th>Observation/Predictor Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,462 (economics)</td>
<td>31</td>
<td>47.16 — largest</td>
</tr>
<tr>
<td></td>
<td>6,559</td>
<td>28</td>
<td>234.3</td>
</tr>
<tr>
<td>3. Antony (1998)</td>
<td>14,837</td>
<td>19</td>
<td>780.89</td>
</tr>
<tr>
<td></td>
<td>12,282</td>
<td>59</td>
<td>208.17 — largest</td>
</tr>
<tr>
<td>5. Bers (1994)</td>
<td>403</td>
<td>6</td>
<td>67.2 — smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>403 — largest</td>
</tr>
<tr>
<td></td>
<td>17,078</td>
<td>47</td>
<td>363.36 — largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>32.5 — largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>40.4 — smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>343.7 — largest</td>
</tr>
<tr>
<td>8. Cabrera, Stampen, &amp; Hansen (1990)</td>
<td>1,375</td>
<td>34</td>
<td>15.25/14.08 — smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>29.55 — largest</td>
</tr>
<tr>
<td></td>
<td>325 (Total)</td>
<td>11</td>
<td>29.55 — largest</td>
</tr>
<tr>
<td></td>
<td>240 (largest)</td>
<td>12</td>
<td>20 — largest</td>
</tr>
<tr>
<td>11. Delucchi (1997)</td>
<td>327</td>
<td>13</td>
<td>25.15</td>
</tr>
<tr>
<td>12. Dey &amp; Astin (1993)</td>
<td>471 (derivation sample)</td>
<td>25</td>
<td>18.84</td>
</tr>
<tr>
<td></td>
<td>476 (confirmation sample)</td>
<td>18</td>
<td>19.04</td>
</tr>
<tr>
<td>13. Dey, Korn, &amp; Sax (1996)</td>
<td>9,402</td>
<td>18</td>
<td>522.33</td>
</tr>
<tr>
<td>14. Feldman (1993)</td>
<td>1,140</td>
<td>15</td>
<td>76</td>
</tr>
<tr>
<td>15. Flint (1997)</td>
<td>1,117</td>
<td>43</td>
<td>25.98 — smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>124.11 — largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(LR not available)</td>
<td>270.8</td>
</tr>
<tr>
<td></td>
<td>483 (public)</td>
<td></td>
<td>3.57 (M)</td>
</tr>
<tr>
<td></td>
<td>952 (private)</td>
<td></td>
<td>3.33 (F)</td>
</tr>
<tr>
<td>17. Grubb (1989)</td>
<td></td>
<td>14 (M)</td>
<td>3.57 (M)</td>
</tr>
<tr>
<td></td>
<td>5,584 (1990 sample)</td>
<td>4</td>
<td>1,554/1,396 — largest</td>
</tr>
<tr>
<td>19. Hearn (1992)</td>
<td></td>
<td>8 (model 3)</td>
<td>667.5 — smallest</td>
</tr>
<tr>
<td>20. House (1995)</td>
<td></td>
<td>4 (model 1)</td>
<td>1,335 — largest</td>
</tr>
<tr>
<td></td>
<td>76 (Male)</td>
<td>8</td>
<td>9.5 — smallest</td>
</tr>
<tr>
<td></td>
<td>179 (total)</td>
<td>8</td>
<td>22.375 — largest</td>
</tr>
<tr>
<td>21. Hurtado, Inkelas, Briggs, &amp; Rhee (1997)</td>
<td>5,666</td>
<td>31</td>
<td>182.77</td>
</tr>
<tr>
<td>22. Jacobs (1999)</td>
<td>1,759</td>
<td>2</td>
<td>879.5 — smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1,759 — largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 (eq. 1)</td>
<td>142.4 — smallest</td>
</tr>
<tr>
<td>23. Leppel (1993)</td>
<td>712</td>
<td>3</td>
<td>237.33 — largest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.8 — smallest</td>
<td>13.8 — smallest</td>
</tr>
<tr>
<td>24. Lindahl &amp; Winship (1994)</td>
<td>262 ($500+)</td>
<td>19</td>
<td>2,728.8 — largest</td>
</tr>
<tr>
<td></td>
<td>51848 ($50–99)</td>
<td></td>
<td>121.86</td>
</tr>
<tr>
<td>25. Mallette &amp; Cabrera (1991)</td>
<td>853</td>
<td>7</td>
<td>142.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29 (model 1)</td>
<td>828.03</td>
</tr>
<tr>
<td></td>
<td>796</td>
<td>34</td>
<td>23.4 — smallest</td>
</tr>
<tr>
<td>27. Nora, Cabrera, Hagedorn, &amp; Pascarella (1996)</td>
<td>466</td>
<td>13</td>
<td>13.7 — smallest</td>
</tr>
<tr>
<td></td>
<td>1,081</td>
<td>31</td>
<td>31.8 — smallest</td>
</tr>
<tr>
<td></td>
<td>181</td>
<td>53</td>
<td>5.3 — smallest</td>
</tr>
<tr>
<td>28. Okun, Benin, &amp; Brandt-Williams (1996)</td>
<td>652</td>
<td>12</td>
<td>54.3 — smallest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>652 — largest</td>
</tr>
<tr>
<td>Articles</td>
<td>Number of Observations</td>
<td>Number of Predictors</td>
<td>Observation/Predictor Ratio</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------</td>
<td>----------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>29. Okunade &amp; Berl (1997)</td>
<td>546</td>
<td>27 (with sample selectivity correction) 26 (without sample selectivity correction)</td>
<td>20.22 (with) 21 (without)</td>
</tr>
<tr>
<td>30. Ott (1988)</td>
<td>455 (85fall) 501 (84fall)</td>
<td>8</td>
<td>56.88 (85) 63.38 (84)</td>
</tr>
<tr>
<td>31. Rindfuss, Kavee, &amp; Cooksey (1995)</td>
<td>126 (men with BA) 473 (women without BA)</td>
<td>3 or 1</td>
<td>42—smallest 427.67—largest</td>
</tr>
<tr>
<td>32. Sedale (1998)</td>
<td>1,734</td>
<td>30</td>
<td>57.8—smallest 433.5—largest</td>
</tr>
<tr>
<td>33. Smith, Anderson, &amp; Lovrich (1995)</td>
<td>525 (task-based) 539 (role-based)</td>
<td>11</td>
<td>47.73 539</td>
</tr>
<tr>
<td>34. Springer, Palmer, Terenzini, Pascarella, &amp; Nora (1996)</td>
<td>1,061</td>
<td>9</td>
<td>117.89</td>
</tr>
<tr>
<td>35. St. John (1990a)</td>
<td>702</td>
<td>15</td>
<td>46.8—smallest</td>
</tr>
<tr>
<td>36. St. John (1990b)</td>
<td>4,338</td>
<td>16</td>
<td>271.1—largest</td>
</tr>
<tr>
<td>37. St. John (1991)</td>
<td>2,320</td>
<td>18</td>
<td>128.9—smallest</td>
</tr>
<tr>
<td>38. St. John (1999)</td>
<td>4,000</td>
<td>18</td>
<td>222.2—largest</td>
</tr>
<tr>
<td>39. St. John, Andrieu, Oescher, &amp; Starkey (1994)</td>
<td>5,115</td>
<td>3, 8, 9, 12, 13, 14</td>
<td>365.4*</td>
</tr>
<tr>
<td>40. St. John &amp; Noell (1989)</td>
<td>5,768</td>
<td>3, 8, 9, 12, 13</td>
<td>582.2*</td>
</tr>
<tr>
<td>38. St. John (1999)</td>
<td>13,003</td>
<td>25</td>
<td>520.1—smallest</td>
</tr>
<tr>
<td>39. St. John, Andrieu, Oescher, &amp; Starkey (1994)</td>
<td>14,938</td>
<td>25</td>
<td>597.5—largest</td>
</tr>
<tr>
<td>42. St. John &amp; Starkey (1995)</td>
<td>8,237</td>
<td>13</td>
<td>633.6—largest</td>
</tr>
<tr>
<td>43. St. John, Paulsen, &amp; Starkey (1996)</td>
<td>2,986</td>
<td>4.5, 7, 8, 12, 17</td>
<td>176.5—smallest</td>
</tr>
<tr>
<td>44. Stage (1988)</td>
<td>3,755</td>
<td>5 (F)</td>
<td>220.9—largest</td>
</tr>
<tr>
<td>45. Strenta, Elliott, Adair, Matier, &amp; Scott (1994)</td>
<td>2,053</td>
<td>26, 26, 29</td>
<td>78.9, 70.8—smallest</td>
</tr>
<tr>
<td>46. Tinto (1997)</td>
<td>16,241</td>
<td>29, 29, 32</td>
<td>623.5, 506.6—largest</td>
</tr>
<tr>
<td>47. Tornquist &amp; Hoenack (1996)</td>
<td>18,836</td>
<td>17, 23, 31, 34, 38, 39, 40</td>
<td>470.9*</td>
</tr>
<tr>
<td>48. Toutkoushian (1994)</td>
<td>313 (total)</td>
<td>5 (F)</td>
<td>64 (M)</td>
</tr>
<tr>
<td>49. Volkwein &amp; Szelest (1995)</td>
<td>185 (F)</td>
<td>2 (M)</td>
<td>760</td>
</tr>
<tr>
<td>50. Volkwein &amp; Lorang (1996)</td>
<td>128 (M)</td>
<td>7</td>
<td>1773.33*</td>
</tr>
<tr>
<td>51. Volkwein, Szelest, Cabrera, &amp; Napierski-Praecel (1998)</td>
<td>5,320 (total sample)</td>
<td>3</td>
<td>57.4</td>
</tr>
<tr>
<td>52. Weiler (1993)</td>
<td>53208</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>336 (total)</td>
<td>5 (eq. 1)</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>205 (citations-err2,3)</td>
<td>6 (eq. 2) or 7 (eq. 3)</td>
<td>34.17 or 29.29</td>
</tr>
<tr>
<td>54. Wddd (1994)</td>
<td>1,519 (table 3)</td>
<td></td>
<td>303.8</td>
</tr>
<tr>
<td>55. Wddd (1994)</td>
<td>1,525 (footnote 8)</td>
<td>5</td>
<td>305</td>
</tr>
<tr>
<td>56. Volkwein, Szelest, Cabrera, &amp; Napierski-Praecel (1998)</td>
<td>2,662</td>
<td>3.7, 8, 9, 10, 14, 15, 17, 18</td>
<td>147.9</td>
</tr>
<tr>
<td>57. Volkwein &amp; Lorang (1996)</td>
<td>4,007</td>
<td>3</td>
<td>222.6</td>
</tr>
<tr>
<td>58. Volkwein, Szelest, Cabrera, &amp; Napierski-Praecel (1998)</td>
<td>190 (N = 229)</td>
<td>3</td>
<td>63.3</td>
</tr>
<tr>
<td>59. Volkwein, Szelest, Cabrera, &amp; Napierski-Praecel (1998)</td>
<td>162 (Hispanic)</td>
<td>31</td>
<td>5.2—smallest</td>
</tr>
<tr>
<td>60. Volkwein, Szelest, Cabrera, &amp; Napierski-Praecel (1998)</td>
<td>5,010 (white)</td>
<td>31</td>
<td>416.1—largest</td>
</tr>
<tr>
<td></td>
<td>52. Weiler (1993)</td>
<td>120</td>
<td>5.7—smallest</td>
</tr>
<tr>
<td></td>
<td>52. Weiler (1993)</td>
<td>980</td>
<td>54.4—largest</td>
</tr>
</tbody>
</table>

Note: Most studies fit multiple logistic models to data. For these studies, only the largest and the smallest observation to predictor ratios are given in this table. Complete information may be obtained from the first author.

*The entry was based on the number of predictors used in the final step of a sequential model.

*The entry was inferred from the article; the exact value was not stated in the table or in the text.
categorical data modeling, also a conservative ratio recommended by most multivariate statisticians for the stability and the significance tests of ML estimators.

The low ratio of observations to predictors invariably resulted in the instability in the parameter estimates. Fortunately, for studies with low ratios (Adams & Becker, 1990; Cleveland-Innes, 1994; Grubb, 1989; House, 1995; Nora, Cabrera, Hagedorn, and Pascarella, 1996; Volkwein, Szelest, Cabrera, and Napieriski-Prancl, 1998; Weiler, 1993), the standard errors of parameter estimates were presented. Readers are therefore able to make a judgment as to whether the low ratio of observations/predictors had caused a problem in the estimation of parameters, that is, the instability of the logistic model.

3. Was Preliminary Analysis Conducted to Investigate the Need to Transform Measurement Scales?

Seven articles investigated the possibility of transforming predictors in order to improve the model fit (Bruggink and Gambhir, 1996; Delucchi, 1997; Gander, 1999; Okun, Benin, and Brandt-Williams, 1996; Springer, Palmer, Terenzini, Pascarella, and Nora, 1996; Tornquist and Hoenack, 1996; Toutkoushian, 1994). These transformations were in the form of squares, log, quadratic form, converting a continuous predictor (e.g., father’s education) to a categorical variable, and weighted predictor (e.g., weighted distance). According to Hosmer and Lemeshow (1989, pp. 88–91), transforming measurement scales of continuous predictors is part of the model-building strategy. This step ensures that each continuous predictor is on a proper scale that is linearly related to the dependant variable, in logit or probit units.


Four types of modeling techniques are possible in logistic regression: (a) direct modeling, (b) sequential modeling, (c) stepwise modeling, and (d) best $k$-predictors modeling. The first two are controlled and implemented by researchers, whereas the latter two are derived primarily by statistical criteria. Of these four modeling techniques, the direct modeling was used most often by 32 studies (56.14%). The sequential modeling was the next popular technique, employed by 18 studies (31.58%). The remaining studies (7, or 12.28%) performed stepwise modeling. Three studies performed both direct and sequential modeling (Hearn, 1992; Mallette and Cabrera, 1991; St. John, 1999), while two performed
both sequential and stepwise modeling (Ott, 1988; Stage, 1988). Our definition of sequential modeling included both hierarchical and block sequential modeling. Based on this definition, two studies were identified as performing sequential modeling, instead of the stepwise modeling as authors claimed (Nora, Cabrera, Hagedorn, and Pascarella, 1996; Volkwein and Szelest, 1995). None performed the best k-predictors modeling.

It was evident from the review that multiple indices were employed by authors in evaluating models obtained: (a) tests of individual parameters in the model (by 44 articles), (b) tests of the overall model (by 40 articles), (c) validation of predicted probabilities (by 34 articles), and (d) goodness-of-fit statistic (by 30 articles). One article did not report any evidence of the model’s fit. One article reported the range of predictors, another the standard errors of residuals.

In terms of validities of predicted probabilities, authors reported percentages of correct classifications, Somers’ D statistic, sensitivity, specificity, false positive, false negative, or concordance pairs. It was unclear to us whether (a) all articles used 50% as the cutoff on predicted probabilities, (b) the Somers’ D statistic was directly quoted from SAS or MINITAB output, or (c) sensitivity/specificity/false positive/false negative were computed by refitting a logistic model to the data with N-1 observations. Answers to these questions would provide an appropriate framework for interpreting these indexes. First, the specific cutoff applied to predicted probabilities influences the percentage of correct classifications (Menard, 2000; Soderstrom and Leitner, 1997). For some studies, such as a persistence study in which 90% of students persisted and only 10% dropped out, 50% may not necessarily be the most suitable cutoff to use. Second, the Somers’ D statistic reported by SAS and MINITAB is inappropriate as a goodness-of-fit index; it is suitable for model comparisons only (Peng and So, 2002). Finally, measures of association, such as sensitivity, specificity, false positive, and false negative, are positively biased if they are computed for the observation which itself was included in deriving the model (Peng and So, 2002).

The goodness-of-fit statistic was presented in terms of chi-square tests or the $R^2$ type of indexes. As stated earlier, the chi-square test based on Hosmer and Lemeshow statistic is both conservative and sensitive to the way in which predicted probabilities are grouped (Hosmer and Lemeshow, 1989, pp. 143–144; Ryan, 1997, p. 279). Deviance-based chi-square tests are correct only if they are based on the difference between two deviances rather than a deviance alone (McCullagh and Nelder, 1989). Further, deviance-based chi-square tests should be computed from covariate patterns. For this reason, the deviance-based chi-square tests computed by SPSS or SYSTAT are incorrect (Peng and So, 1999, 2002). A few authors cited 2.5 as a minimum criterion for the ratio of deviance divided by degrees of freedom, to indicate a good fit of the model to data. Since this rule of thumb was proposed by Stage (1990) for structural equating models.
it needs to be substantiated in empirical studies before it is extended to logistic models.

As for the statistical package employed by authors to derive logistic regression models, four mentioned SAS (Bivin and Rooney, 1999; Hearn, 1992; Windahl and Winship, 1994; Tinto, 1997), two cited SPSS or SPSS-X (Flint, 1997; Smith, Anderson, and Lovrich, 1995), one used BMDP (Ott, 1988), and another GLIM 3.0 (Cabrera, Stampen, and Hansen, 1990). LIMDEP was used by one (Okunade and Berl, 1997) for sample selection correction. The remaining 43 studies did not reference the statistical software actually used. Although this information may appear trivial to readers, two recent studies by Peng and So (1999, 2002) have found that six major statistical packages (SAS, SPSS, BMDP, SYSTAT, MINITAB, and STATA) differed in their treatment of modeling strategies. Peng and So (in press) recommended that a researcher should always check default settings of a logistic regression procedure. Variations in results obtained from various statistical packages may be due to different default settings. According to Peng and So’s review, none of the six statistical packages was found to be completely free of errors.

5. Was Any Possible Interaction or Confounding Examined?

Seventeen out of 52 articles examined the possibility of interaction and/or confounding effects (Bers, 1994; Bivin and Rooney, 1999; Bruggink and Gambhir, 1996; Cabrera, Stampen, and Hansen, 1990; Leppel, 1993.; Lindahl and Winship, 1994; Nora, Cabrera, Hagedorn, and Pascarella, 1996; Okun, Benin, and Brandt-Williams, 1996; Ott, 1988; Rindfuss, Kavee, and Cooksey, 1995; St. John, Paulsen, and Starkey, 1996; St. John and Starkey, 1995; St. John, Andrieu, Oescher, and Starkey, 1994; Stage, 1988; Volkwein, Szelest, Cabrera, and Napierski-Pranci, 1998; Weiler, 1993). Most of the interaction effects investigated involved categorical predictors, such as gender by race. Testing the interaction effect was often performed by (a) separate group analyses, say, applying logistic regression to each gender by race subgroup, (b) sequential modeling, or (c) hierarchical modeling in which interaction effects were investigated only after the corresponding main effects had been tested to be significant. These are acceptable ways of dealing with interaction effects (Peng and So, 2002). Few studies incorporated interactions between continuous predictors into the logistic model.

Although 17 articles claimed that the author(s) investigated the interaction effect, 2 articles (St. John, Andrieu, Oescher, and Starkey, 1994; St. John, Paulsen, and Starkey, 1996) were in fact examining the confounding effect (Hosmer and Lemeshow, 1989, pp. 63–68). Both studies conducted sequential logistic regression modeling in which the magnitude of slope coefficients changed between steps. These changes were evidently the result of the presence of various predic-
tors in the model, instead of the multiplicative effect of two or more predictors. For this reason, it is better to refer to these effects as confounding effects.

6. Was Diagnostic Analysis of Outliers Performed? If so, What Type of Data Was Used?

Diagnostic statistics in logistic regression are derived from covariate patterns in the data. In all of the studies we reviewed, none performed diagnostic analysis or relied on covariate patterns to derive the logistic regression result. Thus, it remains unclear if and how readers of these articles can be confident that these models fit all data used in these studies.

7. How Were the Results of Logistic Regression Reported and Interpreted?

The results of logistic regression were presented in five different yet related formats: the logistic regression equation (the most popular format by 36 articles, or 69.23%), the delta-\(P\) statistic by 20 articles (38.46%), predicted probabilities or marginal probability by 9 articles (17.31%), odds ratio by 5 articles (9.62%), and logit by 3 articles (5.77%). None used relative risk or the kappa index that are more popular in other fields of study, such as health-related fields. Most articles used more than one expression to report their findings.

The logistic regression equation reported is similar to Equation 4 in which logits of the dependent variable were linearly related to a set of predictors. Six out of 36 articles did not include standard errors of slope coefficients. Two out of the six articles plus another four, by different authors, did not include the \(Y\)-intercept for the unstandardized logistic regression equation.\(^1\) The absence of the standard error for the parameter estimate means that readers will not be able to assess the stability of these parameter estimates. The instability of parameter estimates can be caused by the low observation to predictor ratio, as we addressed previously. Missing this piece of information undermined the utility of the logistic regression results.

Similarly, by not including the \(Y\)-intercept in an unstandardized equation, the authors of six articles offered incomplete information. Consequently, readers will not be able to revalidate the result with another sample or at another time/place. This, too, undermined the utility of logistic regression findings.

The delta-\(P\) statistic was the second most used statistic in reporting logistic regression results. As Equation 5 demonstrates, the relationship is not linear between the probability of observing the outcome of interest and the predictors. It is therefore incorrect to convert a regression coefficient, \(\beta_0\) to delta-\(P\) and then interpret the delta-\(P\) value without regard to the value of \(L_0\) (in the case of a continuous predictor) or the reference category (in the case of a categorical
predictor). Unfortunately, in 15 out of 18 (83.33%) articles that reported delta-
P's, the interpretation of these delta-P's was made in isolation of either \( L_o \) or the
reference category. Furthermore, delta-\( P \) was often discussed as an absolute
change in probability given 1 unit-change in a predictor, \( X_j \). This interpretation
of delta-\( P \) is faulty because delta-\( P \) is conditioned on the entire logistic regres-
sion model, hence, other predictors besides \( X_j \) jointly contribute to the slope
coefficient of \( \beta_j \), from which delta-\( P \) is derived.\(^5\) Any discussion of the signifi-
cance of delta-\( P \) should have included the entire logistic regression model and
the reference category or \( L_o \) of \( X_j \).

The third most frequently used form of presenting logistic regression results
was predicted probability or marginal probability. It was reported by 9 out of
52 articles (17.31%). The predicted probability was computed from Equation 4
or 5 for individuals or subsamples with certain characteristics of the predictors.
These individuals or subsamples may be hypothetical or real.\(^3\) From an applica-
tion point of view, these illustrations are an effective mechanism of presenting
logistic regression results. Not only do they demonstrate the utility of logistic
regression models, but also the relative impact of predictors on the likelihood
of the outcome of interest. Since predicted probabilities can be derived from
either Equation 4 or 5, authors should be careful in stating which equation is
used in the calculation and whether the predicted value is the probability (based
on Equation 5), the logit (based on Equation 4), or odds (the exponentiation of
logit, indirectly based on Equation 4). Our review found more than one article
mixed the concept of probability with that of logit, odds, or odds ratio. These
terms are not interchangeable, though they are related conceptually. Marginal
probability (also called marginal effect, partial effect, or partial change) was
defined at least in two different ways in the articles we reviewed.\(^4\) As explained
previously, the concept of marginal probability is not useful for explaining logis-
tic regression models.

Finally, three articles reported logistic regression results on the logit metric.
The logit score is directly predicted from Equation 4 and is linearly related to
all predictors. The difficulty with interpreting logits is that they are not easily
understood and they do not have variance or standard deviation equal to 1 (as
discussed earlier under “Logistic Regression Models”).

8. Was a Multivariate Statistics Reference Cited? If so, What Was It?

One indicator of guidelines that governed the implementation of logistic re-
gression ought to be a multivariate citation among the references. Although the
mere citation of a multivariate reference did not guarantee that such a reference
was the exclusive foundation for the implementation of logistic regression, it
nonetheless provided a potential indicator of what might have been used as a
guideline for the analysis.
In 12 of 52 studies we reviewed, not a single multivariate reference was cited. Among the remaining 40 studies, both multivariate statistics textbooks/software manuals and methodology papers were cited. Table 2 summarizes the frequency with which these references were made.

Among the multivariate textbooks, three references were cited most frequently: Aldrich and Nelson (1984), Fienberg (1983 or earlier editions) and Hanushek and Jackson (1977). These highly appraised textbooks provide reputable information on logistic regression. Five papers published by higher education researchers (Cabrera, 1994, St. John, Kirshstein, and Noell, 1991; Stage, 1990;

<table>
<thead>
<tr>
<th>References—Books and Software Manual</th>
<th>Frequency of Citations</th>
<th>References—Methodology Papers</th>
<th>Frequency of Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanushek &amp; Jackson (1977)</td>
<td>7</td>
<td>Cabrera (1994)</td>
<td>7</td>
</tr>
<tr>
<td>Fienberg (1983 or earlier editions)</td>
<td>7</td>
<td>Press &amp; Wilson (1978)</td>
<td>4</td>
</tr>
<tr>
<td>Pedhazur (1982)</td>
<td>4</td>
<td>Stage (1990)</td>
<td>2</td>
</tr>
<tr>
<td>Maddala (1987, or earlier editions)</td>
<td>3</td>
<td>Amemiya (1984)</td>
<td>2</td>
</tr>
<tr>
<td>Norušis (1990)</td>
<td>2</td>
<td>Becker &amp; Kennedy (1992)</td>
<td>1</td>
</tr>
<tr>
<td>Afifi &amp; Clark (1990)</td>
<td>1</td>
<td>Cleary &amp; Angel (1984)</td>
<td>1</td>
</tr>
<tr>
<td>Agresti (1990)</td>
<td>1</td>
<td>Harrell &amp; Lee (1985)</td>
<td>1</td>
</tr>
<tr>
<td>Backer &amp; Nelder (1988)</td>
<td>1</td>
<td>Heckman (1979)</td>
<td>1</td>
</tr>
<tr>
<td>Davison &amp; MacKinnon (1993)</td>
<td>1</td>
<td>Jackson (1981)</td>
<td>1</td>
</tr>
<tr>
<td>Greene (1989)</td>
<td>1</td>
<td>Kass (1980)</td>
<td>1</td>
</tr>
<tr>
<td>Harrell (1986)</td>
<td>1</td>
<td>Kennedy, P. E. (1981)</td>
<td>1</td>
</tr>
<tr>
<td>Judge, Griffiths, Hill, &amp; Lee (1980)</td>
<td>1</td>
<td>McCullagh (1980)</td>
<td>1</td>
</tr>
<tr>
<td>Judge, Hill, Griffiths, Luptkepohl, &amp; Lee (1982)</td>
<td>1</td>
<td>St. John, Kirshstein, &amp; Noell (1991)</td>
<td>1</td>
</tr>
<tr>
<td>Kennedy, J. J. (1992)</td>
<td>1</td>
<td>Tobin (1958)</td>
<td>1</td>
</tr>
<tr>
<td>Maddala (1992)</td>
<td>1</td>
<td>Weiler (1987)</td>
<td>1</td>
</tr>
<tr>
<td>McCullagh &amp; Nelder (1989)</td>
<td>1</td>
<td>Weiler (1989)</td>
<td>1</td>
</tr>
<tr>
<td>Press (1972)</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tabachnick &amp; Fidell (1983)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Weiler, 1987, 1989) also contributed significantly to the advancement of logistic regression in higher education research. The article by Peterson (1985) was used mainly to explain the use and interpretation of delta-$P$.

SUMMARY AND RECOMMENDATIONS

The review of 52 articles revealed an increasingly sophisticated use and interpretation of logistic regression. Topics researched by logistic regression were diverse, interesting, and meaningful. A summary of our findings is as follows:

1. Logistic regression was conducted in several studies with sample sizes that may not have achieved a reasonable level of stability. Although differences in opinions by multivariate statisticians do not allow one to state a precise standard for the observation/predictor ratio, seven studies (13.46%) had low observation to predictor ratios—lower than standards recommended by at least four textbooks.

2. Seven studies explored the need to transform scales of predictors in order to improve the fit of a model. In several studies, the construct underlying a predictor was measured by a single item on a 5-point Likert scale. Such measurements lack the variability to relate a predictor to the outcome variable, whether the latter is expressed in probabilities or logits.

3. Categorical predictors were typically recoded into a set of dummy variables before they entered logistic regression models. All studies but one transformed categorical predictors correctly. A majority of these studies labeled the reference category or group clearly so that readers may correctly interpret regression coefficients of the dummy variables.

4. More than half of the studies applied direct modeling to derive logistic models. Less than one eighth of studies performed stepwise logistic modeling—a procedure notorious for capitalizing on random errors in data and for which larger samples are required.

5. All studies but one assessed the adequacy of logistic regression models in terms of statistical tests of individual parameters, of the overall model, of the goodness-of-fit statistic, or validities of predicted probabilities. Most studies provided multiple statistics as the evidence of models' fit. However, several of these statistics were calculated incorrectly by commercially available statistical packages. Information on statistical packages was disclosed only in a handful of studies. One study crossvalidated the logistic regression result; none performed the diagnostic analysis.

6. Interactions were typically handled by subgroup analyses when, in fact, they could have been directly specified in the logistic regression model by multiplicative terms.

7. Several studies employed data that contained extremely large or small base-
line proportions of the outcome of interest, such as 96.43% or 1.6%. This in and of itself is not a problem. But an extreme baseline proportion prevents researchers from performing the diagnostic analysis. In this situation, a researcher may wish to consider performing Poisson logistic regression instead (Flury, 1997, pp. 557–558).

8. Selection bias, incurred during the sampling process, was frequently corrected by authors through computer algorithms. Volunteer samples or self-selected samples were usually not corrected. These samples limit the generalizability of regression results.

9. Approximately 70% of articles reported linear logistic regression equations as results. Of these, a few studies did not include the Y-intercept or the standard error of estimated regression coefficients; 38.46% of articles reported the delta-P—a statistic whose interpretation needs to be qualified by its limitations. Likewise, the reporting of marginal probabilities, though by a few studies, lacked clarity and purpose. In quite a few studies, terms such as odds ratio, odds, logit (or log odds), and probability were used in defiance of standard definitions.

In light of our review regarding the use and interpretation of logistic regression, leading research journals in higher education are in a position to help guide researchers in using logistic regression techniques. With the wide availability of sophisticated statistical software installed on high-speed computers, the anticipated use of this technique appears to be increasing. That potential expanded usage can move higher education research toward the elimination of future misuses by bringing clarity to the reporting of results and by avoiding the pitfalls that have been identified. The following recommendations emerged from our review:

1. Although uniform consistency in the use of terminology seems unrealistic, terms should be used along with brief definitions. It is advisable to adhere strictly to the mathematical definitions of standard terms commonly used in logistic regression. These standard terms include probability, marginal probability (or marginal effect or partial effect or partial change), delta-P, odds, odds ratio, and logit (same as log odds). Other terms less commonly used, such as $R^2$, need to be accompanied by brief descriptions. This step will serve to enhance the understanding of results.

2. A clear identification of the outcome of interest, such as persisting in a college (as opposed to dropping out), is not only desirable but essential in understanding logistic regression results. Likewise, the reference category of a categorical predictor needs to be identified.

3. As stated earlier, the marginal probability does not indicate the change in probability for a unit change in $X_i$ (Long, 1997, p. 135). Furthermore, the
sign of the marginal probability is not necessarily the same as the sign of the logistic regression coefficient. It is even possible for the marginal probability of \( X_i \) to change signs as \( X_i \) changes. Only under the condition that a predictor's value varies over an area of the probability curve that is nearly linear, can the marginal probability be used to summarize the effect of a unit change in that predictor on the probability of an outcome. When the probability curve is changing rapidly or when a predictor is a dummy variable, the use of marginal probabilities is misleading. For these reasons, the reporting of marginal probability (also called marginal effect, partial effect, or partial change) should be discouraged.

4. The presentation of delta-\( P \)s should always be accompanied by three additional specifications: (a) the amount of change in a predictor, (b) the starting value of that predictor, and (c) the specific values of all other predictors in the model. Authors should make every effort to avoid the impression that the delta-\( P \) implies a linear association. Nor should delta-\( P \)s be compared across different predictors. For these reasons, the delta-\( P \) is not a summary statistic about a predictor, but rather a numerical index describing the change in probabilities under prespecified conditions.

5. When reporting logistic regression results, authors should include the entire regression equation, including its intercept and the standard error of each regression coefficient. It is recommended that results be accompanied by as complete evaluative statistics as possible. These may include the overall model significance test, the significance test of each regression coefficient, the goodness-of-fit statistics, evidence of the predictive power of the model (such as the percentage of correct classifications), and the diagnostic analysis.

6. The stepwise logistic regression technique should be used as an exploratory tool for the identification of plausible models, rather than for the confirmation or refutation of a specific model for the data.

7. Since logistic regression is a multivariate technique, the interpretation of results can easily capitalize on the data structure specific to the sample used. Therefore, researchers should use large samples and crossvalidate results using procedures such as the jackknifing.

8. Authors should specify statistical software that was used to perform logistic regression. This information can help readers verify statistical indexes calculated and inform readers of programming mistakes and limitations associated with the statistical software actually used.

Implications from the present investigation are multifold; first, readers need to have a good grasp of fundamentals of logistic regression before they are able to discern the wheat from the chaff. Second, standards are needed in editorial policies regarding the appropriate application of logistic regression and regarding the consistency in the Results section. Third, researchers are encouraged to
use this modeling technique more fully in the interpretation of their results so that comparisons can be made across studies. It is our hope that this review has helped to establish new guidelines for journal editors and researchers in formulating editorial policies and practices regarding the use of the versatile logistic regression technique and the communication of its results with readers. These guidelines represent a potentially important step forward in the interpretation and presentation of logistic regression results in higher education research. While prior studies have not always followed these guidelines, all authors are credited for making substantive contributions, as well as for introducing logistic regression into the field. The quality of higher education studies would be enhanced if these new guidelines were widely adopted. At the very least, they would improve comparability across studies that used logistic regression models.

Acknowledgments. Authors wish to thank two anonymous Consulting Editors, Dan J. Mueller, and Lisa Kurz for their constructive comments on an earlier version of this article.

APPENDIX: LIST OF ARTICLES REVIEWED


ENDNOTES

1. Specific evaluative information on each article may be obtained from the first author.

2. Concerning the use of delta-P, Long (1997) rightly points out that "The amount of discrete change in the probability for a change in $X_i$, i.e., delta-P, depends on (1) the amount of change in $X_i$, (2) the starting value of $X_i$, and (3) the values of all other predictors in the model. For example, if we have two predictors, $X_1$ and $X_2$, the change in Probability ($Y = 1 \mid X_1, X_2$) when $X_1$ changes from 1 to 2 does not necessarily equal the change [in probability] when $X_1$ goes from 2 to 3. . . . Moreover, the change in $Pr(Y = 1 \mid X_1, X_2)$ when $X_1$ changes from 1 to 2 with $X_2 = 1$ does not necessarily equal the change [in probability] when $X_2 = 2$. Thus, the practical problem is choosing which values of the variables to consider and how much to let them change" (p. 76).

3. One illustration of real cases was found in Stage (1989) in which two female students' profiles of ethnicity, academic integration score, social integration score, and institutional commitment were first presented. Their predicted probabilities (which should have been odds) were then predicted using the logistic regression equation formulated in the study. Finally, the actual outcome (persisting versus dropping out from the university) was compared with the predicted probabilities. Another illustration of the predicted probabilities for hypothetical individuals was found in Leppel (1993). These hypothetical individuals were supposed to have achieved 1,000 points on SAT and lived either within 10 miles, 10–50 miles, or more than 50 miles from the university which they had been accepted into. Their predicted probabilities of ultimate attendance at that university were calculated from a logistic regression model and shown to be inversely related to the distance from that school, according to a gravity model.

4. According to Sedia (1998, p. 352, footnote a of Table 6), "For each regressor, except for the dummy variables, marginal probabilities were calculated as the effect of one standard deviation increase in the regressor above its sample mean while keeping all other regressors at their respective sample averages. For the dummy variables, marginal effects were calculated as the change in the probability of each outcome when the dummy variable takes each of its two values while keeping all other variables at their respective sample means." By Okunade and Berl's definition (1997, p. 208, footnote b of Table 2), marginal probability is "computed as the logistic distribution density function value (evaluated at the means of the regressor variables) multiplied by the parameter estimate."

REFERENCES


USE AND INTERPRETATION OF LOGISTIC REGRESSION


Received November 13, 2000.