August, 2004

A Comparison of a Logistic Regression Model, a Linear Probability Model, and a Third-Degree Polynomial Model: Which Should a Researcher Use

John Fraas

A Comparison of a Logistic Regression Model, a Linear Probability Model, and a Third-Degree Polynomial Model: Which Should a Researcher Use

Isadore Newman, Ph.D., The University of Akron

Russell Brown, Ph.D., Cleveland State University

John Fraas, Ph.D., Ashland University

Presented at the 2004 annual meeting of the American Educational Research Association in San Diego, California

Introduction

The purpose of this study is to present a comparison of three types of regression models that could be used to analyze dichotomous criterion variables under three different data structures, and to discuss the implications of the results of those comparisons. The three types of models used were (a) logistic regression, (b) linear ordinary least squares, and (c) polynomial ordinary least squares models.

Logistic regression is commonly used in medical literature as a means to account for the variance in a binary (or categorical) dependent variable (King & Ryan, 2002), and its use is growing in the social sciences literature as well. Peng, So, Stage, and St. John (2002) reported that "research using logistic regression has been published with increasing frequency in three higher education journals: *Research in Higher Education, The Review of Higher Education,* and *The Journal of Higher Education*" (p. 259). This trend has corresponded with the increased availability of computer software that provides the option to analyze data using logistic regression (Peng, Lee, & Ingersoll, 2002). While there has been an increase in the use of this method, its use has been accompanied by "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" (Peng, So, Stage, & St. John, 2002, p. 259).

The popularity of logistic regression has grown, in part, due to proponents who have suggested that it is a more appropriate alternative to ordinary least square (OLS) linear regression or discriminant analysis for modeling categorical (dichotomous) dependent variables. With a dichotomous dependent variable, all of the observed

dependent data points will fall on one of two horizontal lines that are parallel, which is a difficult condition to model with the single straight line produced by an OLS linear model. Peng, Lee, and Ingersoll (2002) suggested a potential solution to this problem via plotting the calculated means of the dependent variables for categories of the independent variable. Such a plot takes a sigmoid shape, which Peng, Lee and Ingersoll rightly point out, has extremes that do not follow a linear trend.

Perhaps, it should be no surprise then that a linear fit to such data has obvious limitations. In addition, the errors in this type of model are not normally distributed and are not constant across the range of the data. Finally, OLS models produce values that are above (greater than 1) and below (less than 0) the range of the observed levels of the dependent variable. This problem has been partially addressed by constraining the results of the predicted probabilities to a logical range, but this comes at the expense of treating values above and below the range as perfectly representative of the end points (i.e., 100% likely for points above 1 and 0% likely for points below 0).

The growth in the use of logistic methods is predicated on the reported "superiority of logistic regression over OLS models" (Peng, So, Stage, & St. John, 2002, p. 260) as a means to overcome the "limitations of ordinary least squares (OLS) regression in handling dichotomous outcomes" (Peng & So, 2002, p. 31). The logistic model is as follows:

$$\text{Ln}[p/1\text{-}p] = \alpha + \beta x + e$$

Where: $p =$ probability that the event Y occurs

$\alpha =$ the Y intercept

$\beta =$ the regression coefficient

$e =$ error

The natural log transformation of the odds ratio is necessary to make this relationship linear. The most obvious advantage of this model is that it constrains the predicted values of probability to the logical range of 0 to 1, which overcomes an obvious limitation of the OLS model. Additionally, the model does not require "data that are drawn from a multivariate normal distribution with equal variances and covariances for all the variables" (Peng & So, 2002); and, therefore, it has less restrictive assumptions than either OLS or linear discriminant function analysis.

Brown and Newman (2002) examined methods for modeling data that conformed to a known function or shape and found that, in some instances, polynomial modeling could be superior to modeling based upon the known function (e.g., cosine). A sigmoid curve could be modeled using a polynomial function that accounted for the two inflection points in the curve:

$$Y = a_0U + a_1X + a_2X^2 + a_3X^3 + \text{Error}$$

A polynomial model of this nature would potentially better fit the shape of the distribution at its extremes; and, if consistent with the research of Brown and Newman (2002), it would be better able to fit the data if the data deviated from the sigmoid shape.

While many have expressed concerns about the limitations of OLS when predicting dichotomous outcomes, Pohlmann and Leitner (2000) found very similar results when they compared OLS and logistic regression methods. Indeed, they found identical conclusions in regard to significance testing, and they found the predicted values

Linear, polynomial, and logistic regression analyses were performed on each of the three data sets, and the three methods were compared in terms of (a) the goodness-of-fit values and the statistical tests of those values, (b) the correlations of the predicted probabilities, (c) the mean square error values, and (d) the accuracy of the classifications of group membership.

Results

The three sets of data were analyzed by using a linear OLS model, a polynomial OLS model, and a logistic model. In order to facilitate the retention of all the variables in the polynomial model (i.e., linear, squared, and cubic variables), the scores were centered before these variables were generated. The results of these analyses are contained in Table 1.

Insert Table 1 about here

Goodness-of-Fit Values

The first comparison of the result of the three models involved the amount of variation in the dependent variable accounted for by each model. One issue that had to be addressed before such a comparison could be made is: What value from the logistic regression analysis would be appropriate to compare to the coefficient of determination (R^2) values obtained for the linear and polynomial OLS models. Menard (2000, p. 24) indicated "[the] R^2_L [Cox-Snell R^2 value] has the most intuitively reasonable interpretation as a proportional reduction in error measure, parallel to R^2_O [coefficient of

determination value used in OLS] analogs." Thus, to assess the degree of model fit we compared the Cox-Snell values produced for the logistic regression models and the R^2 values produced for the linear and polynomial OLS models.

As can be seen in Table 1, the results of the analyses are very similar. Under the high correlation condition, the goodness-of-fit values of the linear model (.701), the polynomial model (.764), and the logistic model (.660) are similar. The goodness-of-fit values were also similar for the three models under the medium correlation condition. Specifically the goodness-of-fit values for the linear, polynomial, and logistic models were .198, .199, and .196, respectively. The statistical test for each of these goodness-of-fit values was statistically significant at the .001 level, as was the case under the high correlation condition.

Under the low correlation condition similar goodness-of-fit values were obtained for the linear model (.023), polynomial model (.034), and the logistic model (.030). The statistical tests of the goodness-of-fit values for the linear and logistic models were significant at the .05 level. The statistical test of the R^2 value for the polynomial model was not significant, however, at the .05 level ($p = .079$).

Although these estimates are similar, this is perhaps the least desirable manner to compare the modeling methods as the estimates of relationship have different meanings under the OLS and logistic methods. Estimates of probability and errors in estimation of probability may be more adequate methods of comparison for these methods.

Correlations of the Predicted Probabilities

In order to compare predictions of probabilities, the dependent variable was transformed to reflect the observed probability under each of the independent variable conditions. These values formed the dependent variables used with the linear and polynomial models. If the three models were equally effective, they should produce similar estimates of probability, and the correlations between the estimates of probability under each of the conditions should be high. The correlation coefficient values for the three sets of predicted probability values are listed in Table 2.

[Russ, are these the correlations under the high condition? What about the other conditions?]

Insert Table 2 about here

As was expected, the correlation between the estimated probabilities generated for the estimated probability values generated by the logistic and polynomial models was higher ($r = .968, p < .01$) than the correlation between the estimated probabilities generated by the linear and logistic models ($r = .929, p < .01$).

Mean Square Error Values

Although it is clear from these correlations that there is a great deal of similarity in estimates of probability under each of these conditions, some differences emerge when one compares the mean square error values of the three models (see Table 3). As can be seen in Table 3, the models produce similar mean square error values when there is a

normal distribution with a medium or low degree of relationship between the independent and dependent variables. Differences existed, however, the mean square error terms when a larger relationship existed between the independent and dependent variables and the independent variable had a bimodal distribution.

Insert Table 3 about here

Accuracy of the Classifications of Group Membership

Accuracy of the group membership classifications was the last method used to compare the three types of models. The results of the group classifications for each model under each condition are listed in Table 4. Although the predicted probabilities produced by the various models were not exactly the same, each of the models produced the exact same group membership classifications under the high and medium correlation conditions. In the low correlation condition, however, an interesting difference in the classification patterns emerged. Under this condition, the polynomial model had the lowest mean square error value but it had three more classification errors than either the linear model or the logistic model. In addition, the types of errors (false-positive errors and false-negative errors) made by the polynomial model were different from either the linear or logistic model. The polynomial model, under the low correlation condition, made the fewest false positive classifications, but it made substantially more false negative identifications than did either the linear model or the logistic model.

Insert Table 4 about here

The differences in classification of the sample subjects could be taken as an argument against the polynomial function. This would be a valid conclusion if the sample parameters of false-positive and false-negative identifications were representative of the Type I and Type II error rates in the population as a whole. The probabilities of the subjects, who were classified differently, were located around the cut-value of .50. In the case of the polynomial model, the regression line is virtually horizontal at the mid-point, and therefore, has little predictive (discriminant) power (see Figure 4).

Insert Figure 4 about here

Discussion

The following three main issues regarding the method and results of the study need to be addressed: (a) the distributions chosen for the study, (b) the comparisons made, and (c) implications of the results regarding group membership classifications. One could argue that the distributions chosen for the present study were more alike than different. Certainly, more dramatic differences in distribution shape could have been generated. The distributions were generated with the intent of varying the degree of relationship between the independent and dependent variable ($r = .837$ to $r = .150$) and the shape of the independent variable distribution in terms of the number of modes (one

for both modeling methods to be quite similar. Pohlmann and Leitner did, however, find a slight advantage in accuracy of the predicted values. The similarity in results is striking given that Pohlmann and Leitner used a linear model (OLS) as the basis of the comparison.

Method

The method for the present study was an extension of the method described by Pohlmann and Leitner (2000). Common artificial data sets were used to compare the three methods (i.e., linear, polynomial, and logistic regression) in terms of the effects of known changes in the distribution of the scores on the results of the analysis.

Each of the distributions of the independent variable was created to appear like a group of 200 intelligence test scores. In the first case, which is shown in Figure 2, the scores, which reflected a bimodal distribution (Modes = 94 and 103, M = 98.5, s = 3.154, Range = 15), were highly correlated with the dependent variable (r = .837). In the second case, which is shown in Figure 3, the scores were more normally distributed (M = 97.9, s = 5.636, Range = 26) and moderately correlated (r = .445) with the dependent variable. In scores in the third case (which is shown in Figure 4), were normally distributed (M = 97.2, s = 5.148, Range = 31) and slightly correlated with the dependent variable (r = .150).

Insert Figures 1, 2, and 3 about here

---

or two) and variability around the mean ($s = 3.154$ to $s = 5.636$). Given the similarities in the distribution, one might have expected very similar patterns of outcomes in the comparison of the different modeling techniques, yet, this was not the case.

An examination of the results produced by the three types of models revealed some interesting similarities and differences. Initially, we had posited that the third-degree polynomial and logistic methods would produce more comparable results because the third-degree polynomial modeling would allow the regression line to take a sigmoid shape analogous to that of the logistic model. Brown and Newman (2002) had found that polynomial modeling could, in some instances, be superior to modeling with a known (e.g., cosine) function, and it was expected this could be the case with the logistic model as well. In this study the linear, logistic, and third-degree polynomial modeling methods produced similar goodness-of-fit values. This is not, however, the only means by which the effectiveness of the models can be gauged.

Further comparisons can be made by examining the predicted probabilities and the errors in the predicted probabilities. In this regard, the third-degree polynomial and logistic models produced, as expected, produced more similar results than did the logistic and linear models across each of the three comparison distributions. However, the group classifications produced by these methods did not follow this pattern. Surprisingly, the logistic and linear models produced identical classifications for each of the distribution conditions despite differences in predicted probabilities. The linear, logistic and third-degree polynomial models produced identical classifications in two of the three conditions (high and medium correlations), but in the low correlation condition, the third-

degree polynomial model produced three (3.5%) more errors and produced a different pattern of errors (false-positive and false-negative errors) than did the logistic and linear models.

This last comparison, the pattern of classifications, resulted in the most interesting contrast. Not surprisingly, the cases that were classified differently occurred near the cut-value of the predicted probabilities. This finding points to a concern regarding the stability of the predicted classifications for these modeling techniques. Cases that have predicted probabilities that are above the cut-value (irrespective of differences in the predicted probability) are grouped together as are cases below the cut-value. This aggregation of cases into categories does not take into consideration the error in the predicted probability and, in turn, the stability of the model. The differences in classification produced by the models in the low correlation condition points to the need to develop a method of estimate the stability of these models when using them for classification purposes.

Studies that fail to take this into consideration may unintentionally convey a greater sense of stability in the classifications provided by the models in the study. We would suggest replication as one possible means to estimate the stability of the model (Newman, McNeil, & Fraas, 2003). We strongly believe the argument between statistical significance and practical significance is not as salient as this issue of replicability: and, therefore, we suggest future research may wish to develop methods of estimating the stability (i.e. replicability) of the classifications produced by logistic models. To this end, the authors are presently working on a method of estimating the stability of the predicted

probabilities using confidence intervals around the predicted scores and comparing the stability estimates across the three methods described in this paper.

References

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

King, J. (2002). *Logistic regression: Going beyond point-and-click.* Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).

Menard, S. (1995). *Applied logistic regression analysis.* Thousand Oaks, CA: Sage.

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

Newman, I., McNeil, K., & Fraas, J. (2003). *Deja Vu: Another Call for Replications of Research, Again.* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

Peng, C., Lee, K., & Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*(1), 3-13.

Peng, C., & So, T. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics 1*(1), 31-70.

Peng, C., So, T., Stage, F., St. John, E. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Journal of Research in Higher Education, 43*(3), 259-293.

Pohlmann, J., & Leitner, D. (2004). *A comparison of ordinary least squares and logistic regression.* Manuscript submitted for publication.

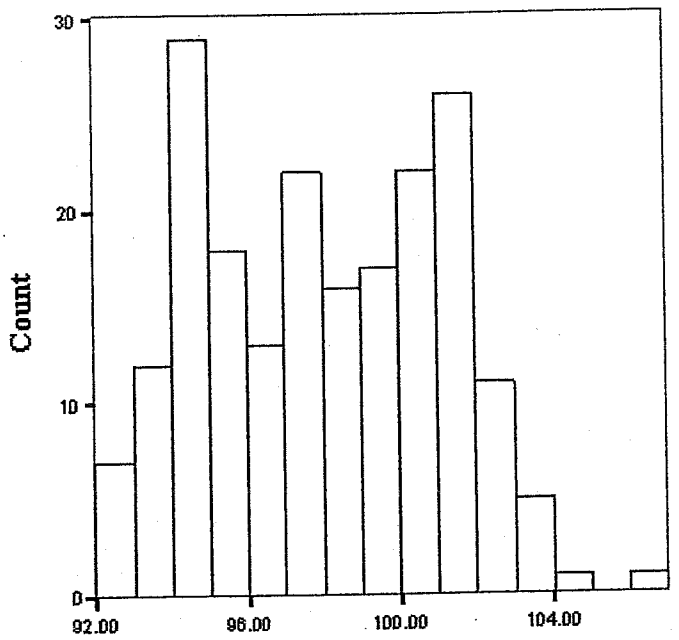

Figure 1. Bimodal distribution with high correlation (r = .837)

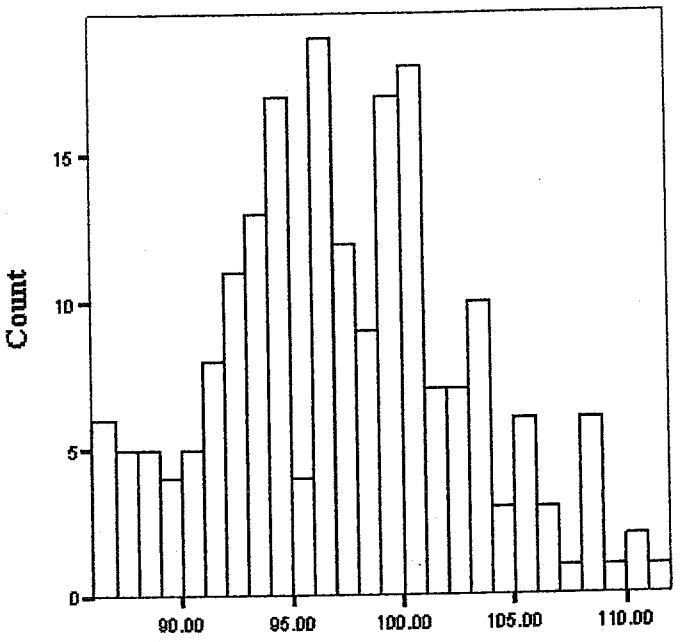

Figure 2. Normal distribution with moderate correlation (r = .445).

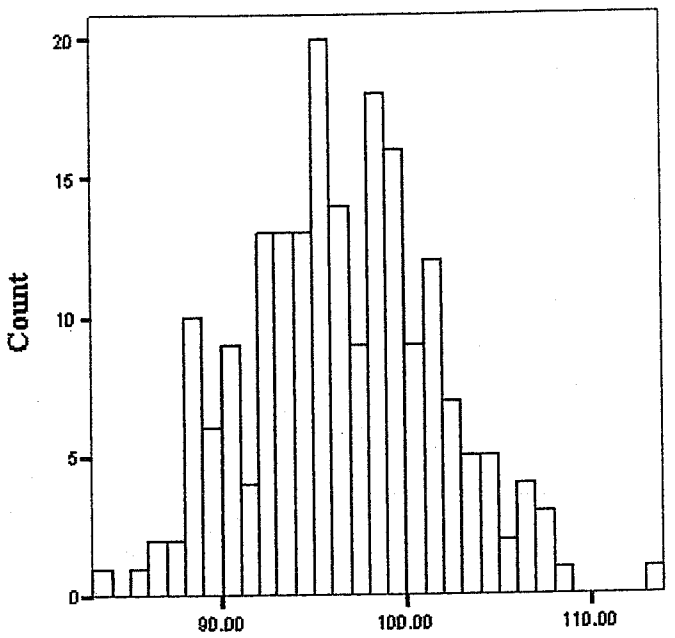

Figure 3. Normal distribution with low correlation (r = .150).

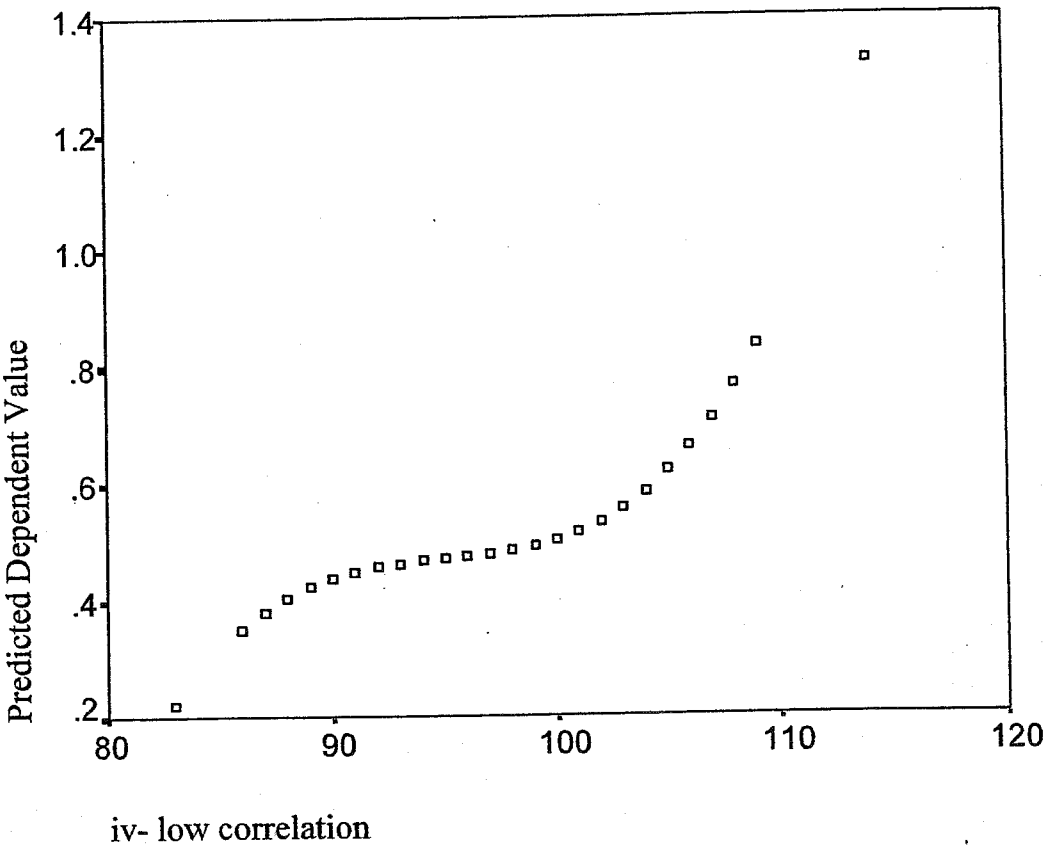

Figure 4. Predicted values of the polynomial model under the low correlation condition.

Table 1

Results of the Comparisons of the Tests of Significance

Model and Condition				
Bimodal Distribution (High Correlation)	*F*	*p*	R^2	Adj. R^2
Linear Model	464.96	<.001	.701	.700
Polynomial (cubic) Model	211.05	<.001	.764	.760
	χ^2	*p*	Cox-Snell	
Logistic Model	15.932	<.001	.660	
Normal Distribution (Med. Correlation)	*F*	*p*	R^2	Adj. R^2
Linear Model	48.80	<.001	.198	.194
Polynomial (cubic) Model	16.27	<.001	.199	.187
	χ^2	*p*	Cox-Snell	
Logistic Model	43.649	<.001	.196	
Normal Distribution (Low Correlation)	*F*	*p*	R^2	Adj. R^2
Linear Model	4.55	.034	.023	.018
Polynomial (cubic)	2.30	.079	.034	.019
	χ^2	*p*	Cox -Snell	
Logistic Model	4.548	.030	.022	

Table 2

Correlations of Predicted Probabilities

Regression Model	Linear	Cubic	Logistic
Linear	-	.941[a]	.929[a]
Cubic		-	.968[a]
Logistic			-

[a] $p < .001$.

Table 3

Mean Square Error Values for Each Condition

Model and Condition	Sum of Squared Error	Mean Square Error
Bimodal Distribution (High Correlation)		
Linear Model	6.17	.0386
Polynomial (cubic) Model	1.82	.0091
Logistic Model	.75	.0038
Normal Distribution (Medium Correlation)		
Linear Model	4.74	.0237
Polynomial (cubic) Model	4.51	.0226
Logistic Model	4.74	.0237
Normal Distribution (Low Correlation)		
Linear Model	5.09	.0255
Polynomial (cubic) Model	4.52	.0226
Logistic Model	5.12	.0256

Table 4

Group Membership Classifications and Errors

Model and Condition	Correct Classification	False Positives	False Negatives
Bimodal Distribution (High Correlation)			
Linear Model	185	7	8
Polynomial (cubic) Model	185	7	8
Logistic Model	185	7	8
Normal Distribution (Medium Correlation)			
Linear Model	135	34	31
Polynomial (cubic) Model	135	34	31
Logistic Model	135	34	31
Normal Distribution (Low Correlation)			
Linear Model	114	39	47
Polynomial (cubic) Model	111	27	62
Logistic Model	114	39	47