A Comparison of Ordinary Least Squares and Logistic Regression[1]

JOHN T. POHLMANN AND DENNIS W. LEITNER, Department of Educational Psychology, Southern Illinois University, Carbondale, IL 62901

ABSTRACT. This paper compares ordinary least squares (OLS) and logistic regression in terms of their underlying assumptions and results obtained on common data sets. Two data sets were analyzed with both methods. In the respective studies, the dependent variables were binary codes of 1) dropping out of school and 2) attending a private college. Results of both analyses were very similar. Significance tests (alpha = 0.05) produced identical decisions. OLS and logistic predicted values were highly correlated. Predicted classifications on the dependent variable were identical in study 1 and very similar in study 2. Logistic regression yielded more accurate predictions of dependent variable probabilities as measured by the average squared differences between the observed and predicted probabilities. It was concluded that both models can be used to test relationships with a binary criterion. However, logistic regression is superior to OLS at predicting the probability of an attribute, and should be the model of choice for that application.

OHIO J SCI 103 (5):118–125, 2003

INTRODUCTION

Logistic regression analysis is one of the most frequently used statistical procedures, and is especially common in medical research (King and Ryan 2002). The technique is becoming more popular in social science research. Ordinary least squares (OLS) regression, in its various forms (correlation, multiple regression, ANOVA), is the most common linear model analysis in the social sciences. OLS models are a standard topic in a one-year social science statistics course and are better known among a wider audience. If a dependent variable is a binary outcome, an analyst can choose among discriminant analysis and OLS, logistic or probit regression. OLS and logistic regression are the most common models used with binary outcomes. This paper compares these two analyses based on their underlying structural assumptions and the results they produce on a common data set.

Logistic regression estimates the probability of an outcome. Events are coded as binary variables with a value of 1 representing the occurrence of a target outcome, and a value of zero representing its absence. OLS can also model binary variables using linear probability models (Menard 1995, p 6). OLS may give predicted values beyond the range (0,1), but the analysis may still be useful for classification and hypothesis testing. The normal distribution and homogeneous error variance assumptions of OLS will likely be violated with a binary dependent variable, especially when the probability of the dependent event varies widely. Both models allow continuous, ordinal and/or categorical independent variables.

Logistic regression models estimate probabilities of events as functions of independent variables. Let y_i represent a value on the dependent variable for case i, and the values of k independent variables for this same case be represented as x_{ij} (j = 1,k). Suppose **Y** is a binary variable measuring membership in some group. Coding y_i = 1 if case i is a member of that group and 0 otherwise, then let p_i = the probability that y_i = 1. The odds that y_i = 1 is given by $p_i/(1-p_i)$. The log odds or logit of p_i equals the natural logarithm of $p_i/(1-p_i)$. Logistic regression estimates the log odds as a linear combination of the independent variables:

$$\text{logit}(p) = ß_0 + ß_1\mathbf{X}_1 + ß_2\mathbf{X}_2 + \ldots + ß_k\mathbf{X}_k,$$

where ($ß_0 \ldots ß_k$) are maximum likelihood estimates of the logistic regression coefficients, and the **X**s are column vectors of the values for the independent variables. A coefficient assigned to an independent variable is interpreted as the change in the logit (log odds that y = 1), for a 1-unit increase in the independent variable, with the other independent variables held constant. Unlike the closed form solutions in OLS regression, logistic regression coefficients are estimated iteratively (SAS Institute Inc. 1989). The individual y_i values are assumed to be Bernoulli trials with a probability of success given by the predicted probability from the logistic model.

The logistic regression model predicts logit values for each case as linear combinations of the independent variable values. A predicted logit for case i is obtained from the solved logistic regression equation by substituting the case's values of the independent variables into the sample estimate of the logistics regression equation,

$$\text{logit}_i = b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_mx_{ik}.$$

The predicted probability for case i is then given by

$$p_i = \exp(\text{logit}_i) / [1 + \exp(\text{logit}_i)].$$

This value serves as the Bernoulli parameter for the binomial distribution of **Y** at the values of **X** observed for case i. Logit values can range from minus to plus infinity, and their associated probabilities range from 0 to 1. Tests of significance for the logistic regression coefficients ($ß_i$) are performed most commonly with the Wald χ^2 statistic (Menard 1995, p 39), which is based on the change in the likelihood function when an independent variable is added to the model. The Wald χ^2 serves the same role as the *t* or *F* tests of OLS partial regression coefficients. Various likelihood function statistics are also available to assess goodness of fit (Cox and

[1]Manuscript received 11 November 2002 and in revised form 13 May 2003 (#02-28).

Snell 1989, p 71).

On the other hand, ordinary least squares (OLS) models the relationship between a dependent variable and a collection of independent variables. The value of a dependent variable is defined as a linear combination of the independent variables plus an error term,

$$\mathbf{Y} = ß_0 + ß_1\mathbf{X}_1 + ß_2\mathbf{X}_2 + \cdots + ß_k\mathbf{X}_k + \varepsilon,$$

where the ßs are the regression coefficients, **X**s are column vectors for the independent variables and ε is a vector of errors of prediction. The model is linear in the ß parameters, but may be used to fit nonlinear relationships between the **X**s and **Y**. The regression coefficients are interpreted as the change in the expected value of **Y** associated with a one-unit increase in an independent variable, with the other independent variables held constant. The errors are assumed to be normally distributed with an expected value of zero and a common variance. In a random sample, the model is represented as

$$\mathbf{Y} = b_0 + b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \ldots + b_k\mathbf{X}_k + E,$$

and its coefficients are estimated by least squares; the solution for the weights (b_i) minimizes the sample error sum of squares (**E'E**). Closed-form and unique estimates of least squares coefficients exist if the covariance of **X** is full rank. Otherwise, generalized inverse approaches can produce solutions (Hocking 1985, p 127). Inferential tests are available for the ßs, individually or in combinations. In fact, any linear combination of the ßs can be tested with an *F*-test.

The sample predicted Y values ($\hat{\mathbf{Y}}$) are obtained for case i by substituting the case's values for the independent variables in the sample regression equation:

$$\hat{Y}_i = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_{ik},$$

When **Y** is a binary variable, $\hat{\mathbf{Y}}$ values estimate the probability that $Y_i = 1$. While probabilities range between 0 and 1, OLS predicted Y values might fall outside of the interval (0,1). Out-of-range predictions like this are usually the result of linear extrapolation errors when a relationship is nonlinear. This problem can be solved after the analysis by changing negative predicted values to 0, and values greater than 1 to 1. These adjusted predictions are no longer OLS estimates, but they might be useful estimates of the probability that $y_i = 1$, and are certainly more accurate than out-of-range OLS values.

Major statistical packages such as SAS (SAS Institute Inc. 1989), have excellent software for analyzing OLS and logistic models. The choice of models can be made based on the data and the purpose of the analysis. Statistical models can be used for prediction, classification and/or explanation. This study will assess the relative effectiveness of OLS and logistic regression for these purposes by applying them to common data sets and contrasting the results.

MATERIALS AND METHODS

In order to compare OLS and logistic regression, common data sets were analyzed with both models, and the results were contrasted. Monte-Carlo methods were not used for this comparison because one would have to use either the OLS or logistic structural model to generate the data. The comparative results would certainly favor the generating model. By using two empirical research data sets, no artificially induced structural biases are present in the data.

The first data set was taken from a popular social science statistics text (Howell 2002, p 728). The data set has nine variables measured on 88 high school students. The variables are described in Table 1. The dependent variable (DROPOUT) was a binary variable coded 1 if the student dropped out of school and 0 otherwise. The remaining variables were used as independent variables in this analysis. The second data set was a 600 case extract from the High School and Beyond project (National Center for Education Statistics 1995). The dependent variable was coded 1 if the case attended a private college, 0 otherwise. The independent variables were high school measures of demographics, personality, educational program and achievement scores. The variables are described in Table 2.

Statistical Analysis System (SAS Institute Inc. 1989) programs were used to analyze the data. PROC REG was used to perform the OLS analysis and PROC LOGISTIC was used for the logistic regression model. The binary outcome variables served as the dependent variables and the remaining variables listed in Tables 1 and 2 were respectively used as independent variables. The models were compared by examining four results: 1) model and individual variable significance tests, 2) predicted probabilities that y = 1, 3) accuracy of the estimates of the probability that y = 1, and 4) accuracy of classifications as a member of the group coded as a 1.

Methods — Dropout Study

For the dropout study, predicted values for both models were correlated using Pearson and Spearman correlation coefficients. The Spearman coefficient measures rank correlation and the Pearson correlation measures linear association between the OLS and logistic predicted values. The OLS predicted values were also adjusted to range between (0,1). This adjusted variable is labeled OLS01 in Table 1. There were no OLS predicted values greater than 1 but there were 11 negative values. The predicted values were used to classify cases as a dropout. Ten cases (11%) dropped out of school. A case was classified as a dropout if its predicted value was among the highest ten in the sample.

Lastly, the sample was ranked with respect to the predicted probabilities for OLS and logistic regression. Eight subgroups of the ranked probabilities were formed. Actual probabilities of DROPOUT were calculated in each subgroup and compared to the average OLS and logistic regression probability estimates. Hosmer and Lemeshow (1989) developed a χ^2 goodness-of-fit test for logistic regression by dividing the sample into ten, equal sized ranked categories based on the predicted values from the logistic model and then contrasting frequencies based on predicted probabilities with observed frequencies. The Hosmer and Lemeshow

TABLE 1

OLS and logistic regression solutions for the Howell data (dependent variable = DROPOUT).

	OLS Regression		Logistic Regression	
	Significance Test [a]	*p*-value	SignificanceTest [b]	*p*-value
			Cox and Snell	
	$\mathbf{R}^2 = 0.394$		$\mathbf{R}^2 = 0.287$	
Model Test	$F_{8,79} = 6.431$	<0.0001	$\chi^2 = 34.707$	<0.0001
INTERCPT	0.746	0.4578	0.4442	0.5051
ADDSC [c]	-0.043	0.9659	0.0005	0.9830
GENDER	0.848	0.3990	0.5824	0.4454
REPEAT	4.023	0.0001	6.5712	0.0104
IQ	-0.348	0.7286	0.925	0.3362
ENGL	-0.452	0.6526	0.2857	0.593
ENGG	-0.898	0.3721	0.6394	0.4239
GPA	-0.173	0.8629	0.0772	0.7812
SOCPROB	3.753	0.0003	6.0516	0.0139

[a] Model test is an F statistic with 8,79 d.f. Variable tests are *t* statistics with 79 d.f.

[b] Model test is a Chi-square statistic with 8 d.f. Variable tests are Chi Square statistics with 1 d.f.

[c] Variables are: ADDSC = attention deficit disorder score; GENDER (1 = male 2 = female); REPEAT (1 = repeated a grade in school, 0 = did not repeat a grade); IQ = score on a group IQ test; ENGL = level of English in ninth grade (1 = college prep, 2 = general, 3 = remedial); ENGLG = grades in English; GPA = grade point average; SOCPROB = social problems in ninth grade (1 = yes, 0 = no); DROPOUT (1 = dropped out of school, 0 = did not drop out).

test is not generally applicable to OLS regression because it is possible to obtain negative predicted values and hence, negative frequencies. A simple descriptive measure of accuracy was developed, along the lines of the Hosmer and Lemeshow measure, so OLS and logistic regression solutions could be compared. Assume the cases have been ranked into k ordered categories on the predicted values from a model. Mean square error (MSE) accuracy measures were calculated using the following formula:

$$MSE = \sum_{i=1}^{8} (p_i - \hat{p}_i)^2/k ,$$

where $(p_i - \hat{p}_i)$ is the difference between the observed and average predicted probabilities of being coded 1 on **Y** for the i-th ordered category of predicted values. MSE is a measure of a model's accuracy in estimating the observed probability of an event. If a model gives the same probabilities as the actual probabilities, MSE will be zero. In general, low values of MSE indicate accurate estimation.

Methods — High School And Beyond Study

Essentially the same methods were employed with the High School and Beyond data, except a cross validation study was performed because of the larger sample size. The cross validation was performed by randomly dividing the total sample into two subsamples of size 300. OLS and logistics model solutions were alternately applied from each subsample to the other subsample. For example, the regression coefficients (both OLS and logistic) from subsample 1 were applied to the data of subsample 2. The procedure was then alternated, reversing the roles of the subsamples. Because of the larger sample size , the MSE statistic was calculated on ten ranked categories of the sample rather than the eight categories used in the dropout study.

RESULTS

Dropout Study

Table 1 presents the OLS and logistic regression solutions predicting DROPOUT from the remaining 8 variables. The OLS and logistic regression results are presented for the full model and each independent variable. The regression coefficients are not presented because OLS and logistic coefficients are not directly comparable. OLS coefficients measure changes in expected values of DROPOUT, while logistic coefficients measure changes in the log odds that DROPOUT = 1.

On the other hand, the significance tests are comparable. The null hypotheses for both OLS and logistic regression are identical. That hypothesis states the probability that y = 1 is a constant for all values of the independent variable(s). Graphically, this hypothesis can

TABLE 2

OLS and logistic regression solutions for the High School and Beyond Data dependent variable: school type (1 = private, 0 = public).

	OLS Regression		Logistic Regression	
	Significance Test [a]	*p*-value	Significance Test [b]	*p*-value
			Cox and Snell	
	$\mathbf{R}^2 = 0.089$		$\mathbf{R}^2 = 0.095$	
Model Test	$F_{12,587} = 4.765$	<0.0001	$\chi^2 = 59.823$	<0.0001
SEXM1F2 [c]	0.557(1)	0.456	0.431(1)	0.512
RACE	3.463(3)	0.016	9.198(3)	0.027
SOCSTAT	3.024(2)	0.049	6.189(2)	0.045
HSPROG	14.371(2)	<0.0001	25.395(2)	<0.0001
LOCONTRO	0.024(1)	0.878	0.039(1)	0.844
SELFCONC	0.000(1)	0.988	0.007(1)	0.932
MOTIVAT	0.000(1)	0.996	0.007(1)	0.936
COMPOSIT	1.750(1)	0.186	1.435(1)	0.231

[a] Model test is an F statistic with (12,587) d.f. Variable tests are F statistics with numerator d.f. in parentheses.

[b] Model test is a Chi-square statistic with 12 d.f. Variable tests are Wald Chi Square statistics with d.f. in parentheses.

[c] Variables are: SEXM1F2 (sex: male = 1, female = 2); RACE (1 = Hispanic, 2 = Asian, 3 = African-American, 4 = white); SOCSTAT (social status: 1 = low, 2 = medium, 3 = high); HSPROG (high school program: 1 = college prep, 2 = general, 3 = remedial); LOCONTRO = standardized measure of locus of control; SELFCONC = standardized measure of self-concept; MOTIVAT = measure of motivation (average of three motivational items); COMPOSIT = composite measure of achievement in reading, writing, mathematics, science and social studies.

be graphed as a horizontal regression line. If the same data and same models are used, and the null hypothesis is true in an OLS model, it must be true in the logistic regression model. The test statistics differ between OLS (*t* and *F*) and logistic (χ^2) but they should produce the same rejection decisions. Table 1 shows that all 11 significance tests at alpha = 0.05 agreed between OLS and logistic regression. Both model tests were significant, and the same independent variables (SOCPROB, REPEAT) were significant.

Table 3 presents the results for the predicted probability that a case was a dropout. The first row of Table 1 presents the Pearson correlations between DROPOUT and various estimates of the probability that a student will drop out of school. Logistic regression estimates of the probability of a dropout correlated strongest with DROPOUT (r = 0.694), while OLS predicted values correlated 0.628. The difference between the OLS and Logistic probability correlations with DROPOUT was statistically significant (t = 2.18, df = 85, p <0.05). Logistic and OLS predicted values correlated 0.925 and the Spearman rank correlation between the OLS and logistic estimates is 0.969. The fact that the Spearman rank correlation is larger than the Pearson correlation suggests a nonlinear but monotonic relationship between the OLS and logistic probability estimates.

OLS predicted values ranged from -0.111 to 0.786, with 28 negative values. A new variable was created by modifying the OLS values to conform to the (0,1) interval. This modified variable appears as OLS01 in Table 3 and its values ranged from 0 to 0.786. OLS01 has a slightly higher correlation with DROPOUT (r = 0.636) than the raw OLS values.

These models can also be used for classification. In the present application the predicted values were reduced to a binary variable; 1 = predicted dropout, 0 otherwise. Ten cases actually dropped out of school. In order to match the observed frequencies of DROPOUT = 1 and 0, OLS and logistic predicted probabilities were converted to a binary group prediction by classifying the top 10 predicted values as dropouts and the remaining 78 values as non-dropouts. The OLS and logistic regressions identified the same 10 cases as dropouts. Seven of the 10 cases classified as dropouts were in fact dropouts. Seventy-five of 78 cases were accurately classified as non-dropouts. Overall, 93% of the cases were accurately classified.

Mean square error (MSE) values were computed for logistic, OLS and adjusted OLS (OLS01) predicted values. Table 4 provides an illustration of the MSE calculation for the OLS and logistic models. Because of the small sample size, eight ordered categories of predicted values were used. Logistic regression produced the lowest value (MSE = 0.0010) and therefore was the most accurate model. OLS had the highest value (MSE = 0.0090) and adjusted OLS (OLS01) estimates produced an MSE

TABLE 3

Correlations among predicted probabilities and DROPOUT (Howell data, n = 88).

	1.	2.	3.	4.	5.	6.	7.	8.
1. DROPOUT	1.000	0.694	0.628	0.636	0.467	0.444	0.662	0.662
2. Logistic Reg.		1.000	0.925	0.937	0.702	0.695	0.914	0.914
3. OLS Reg.			1.000	0.994	0.832	0.850	0.777	0.777
4. OLS01				1.000	0.789	0.803	0.791	0.791
5. Ranked Log.					1.000	0.969	0.550	0.550
6. Ranked OLS						1.000	0.550	0.550
7. Binary Log.							1.000	1.000
8. Binary OLS								1.000

Variable Descriptions:
DROPOUT: 1 = dropped out of school, 0 = did not drop; Logistic Reg.: logistic regression predicted prob. (DROPOUT = 1); OLS Reg.: OLS regression predicted value of DROPOUT; OLS01: adjusted OLS predicted values (negative values set to 0); Ranked Log.: ranked values of Logistic Reg.; Ranked OLS: ranked values of OLS Reg.; Binary Log.: top 10 values of Logistic Reg. set to 1,0 otherwise; Binary OLS: top 10 values of OLS Reg. set to 1,0 otherwise.

of 0.0078. All of these MSE values are low, but logistic regression was the most accurate model for predicting the probability that a student will drop out of school.

High School and Beyond Study

Table 2 presents the OLS and logistic regression solutions predicting School Type from the remaining 8 variables. Race, Social Status, and High School Program were multicategory classifications and coded with dummy variables. Only the main effect tests are presented in Table 4. The full model test and all eight partial significance tests at alpha = 0.05 yielded the same results. Both OLY and logistic regression yielded a significant global test; and RACE, SOCSTAT and HSPROG were the significant independent variables. White, upper class students who took a college preparatory program in high school had significantly higher probabilities of attending a private college.

Table 5 presents the results for the predicted probability that a case attended a private college. The first row of Table 5 presents the Pearson correlations between SCHTYPE and various estimates of the probability that a student will attend a private school. Logistic regression predictions of the probability of a dropout correlated strongest with SCHTYPE (r = 0.323), while OLS predicted values correlated 0.298. The difference between the OLS and Logistic probability correlations with SCHTYPE was not statistically significant (t = 0.89, df = 597, p >0.05). Logistic and OLS estimates correlated 0.957, and the rank correlation between the OLS and logistic estimates was 0.992.

OLS predicted values ranged from -0.15 to 0.35, with 47 negative values. A new variable was created by modifying the OLS values to conform to the (0,1) interval. This modified variable appears as OLS01 in Table 5 and its values ranged from 0 to 0.35. OLS01 has a slightly higher correlation with SCHTYPE (r = 0.306) than the raw OLS values.

The predicted values were reduced to a binary variable; 1 = predicted private school, 0 otherwise. Ninety-four cases actually attended a private school. OLS and logistic predicted probabilities were converted to a binary group prediction by classifying the top 94 predicted values as attending a private school and the remaining 503 values as not attending a private school. Logistic regression accurately classified 484 cases (81%). OLS regression accurately classified 480 cases (80%).

Mean square error (MSE) values were computed for logistic, OLS, and adjusted OLS (OLS01) predicted values. Logistic regression produced the lowest value (MSE = 0.0011) and therefore was the most accurate model. OLS had the highest value (MSE = 0.0034) and adjusted OLS (OLS01) estimates produced an MSE of .0029. All these MSE values are low, but logistic regression produced the most accurate model for predicting the probability that a student will attend a private school. Figure 1 provides a graphic representation of this goodness of fit information. Figure 1 shows that logistic regression is superior to OLS in the extreme prediction ranges, especially when the predicted values are close to 0. Figure 2 presents a scatter diagram of the OLS and logistic predicted values. Figure 2 clearly shows how the logistic predicted values are constrained within the (0,1) range, while the OLS estimates can range outside of (0,1).

DISCUSSION

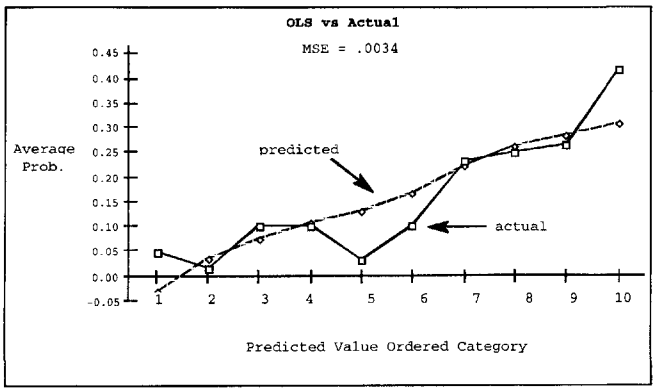
OLS and logistic regression analysis produced very similar results when applied to the same two data sets

TABLE 4

MSE[a] calculations for OLS and logistic regression solutions for the Howell data (n = 88, 11 cases per ordered category).

(i) Predicted Value Ordered Category	Logistic Regression p_i Actual Prob. DROPOUT = 1	Logistic Regression $\hat{p}_i$ Average Predicted Value	OLS Regression p_i Actual Prob. DROPOUT = 1	OLS Regression $\hat{p}_i$ Average Predicted Value
1	0.0000	0.0014	0.0000	-0.0663
2	0.0000	0.0036	0.0000	-0.0241
3	0.0000	0.0067	0.0909	0.0028
4	0.0000	0.0127	0.0000	0.0270
5	0.0909	0.0220	0.0000	0.0495
6	0.0000	0.0540	0.0909	0.0848
7	0.1818	0.1767	0.0909	0.3016
8	0.6363	0.6320	0.6363	0.5338
MSE		0.0010		.0090

[a] $MSE = \sum_{i=1}^{8} (p_i - \hat{p}_i)^2 / 8$



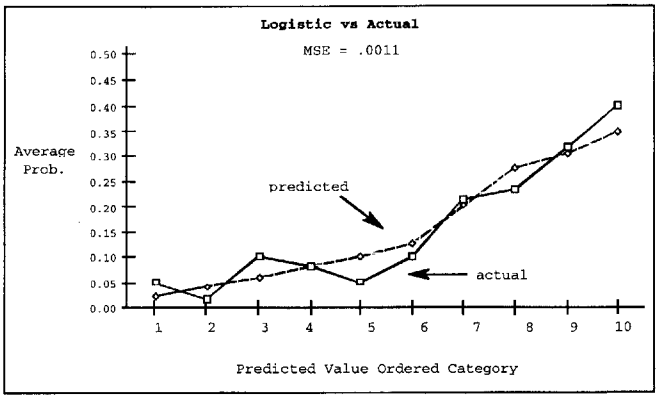


FIGURE 1. Goodness of fit comparison between OLS and logistic regression results for the high school and beyond data.

TABLE 5

Correlations among predicted probabilities and school type for High School and Beyond Data (n = 600).

	1	2	3	4	5	6	7	8	9	10
1. SCHTYPE	1.000	0.323	0.298	0.306	0.268	0.243	0.309	0.302	0.220	0.194
2. Log Reg.		1.000	0.957	0.972	0.676	0.675	0.969	0.964	0.872	0.827
3. OLS Reg.			1.000	0.993	0.575	0.577	0.985	0.991	0.844	0.872
4. OLS_01				1.000	0.594	0.595	0.989	0.995	0.856	0.866
5. BINLOG					1.000	0.924	0.630	0.626	0.573	0.483
6. BINOLS						1.000	0.627	0.630	0.587	0.492
7. RANKLOG							1.000	0.992	0.846	0.851
8. RANKOLS								1.000	0.849	0.863
9. LOGCROSS									1.000	0.929
10. OLSCROS										1.000

Variable Descriptions:

1. SCHTYPE: 1 = attended a private college, 0 = otherwise.
2. Log Reg.: logistic regression predicted prob.(SCHTYPE = 1).
3. OLS Reg.: OLS regression predicted value of SCHTYPE.
4. OLS01: adjusted OLS predicted values (negative values set to 0).
5. BINLOG: top 94 values of Logistic Reg. set to 1,0 otherwise.
6. BINOLS: top 94 values of OLS Reg. set to 1,0 otherwise.
7. RANKLOG: ranked predicted values from Log Reg.
8. RANKOLS: ranked predicted values from OLS Reg.
9. LOGCROSS: cross validated logistic predicted probabilities.
10. OLSCROS: cross validated OLS predicted values.

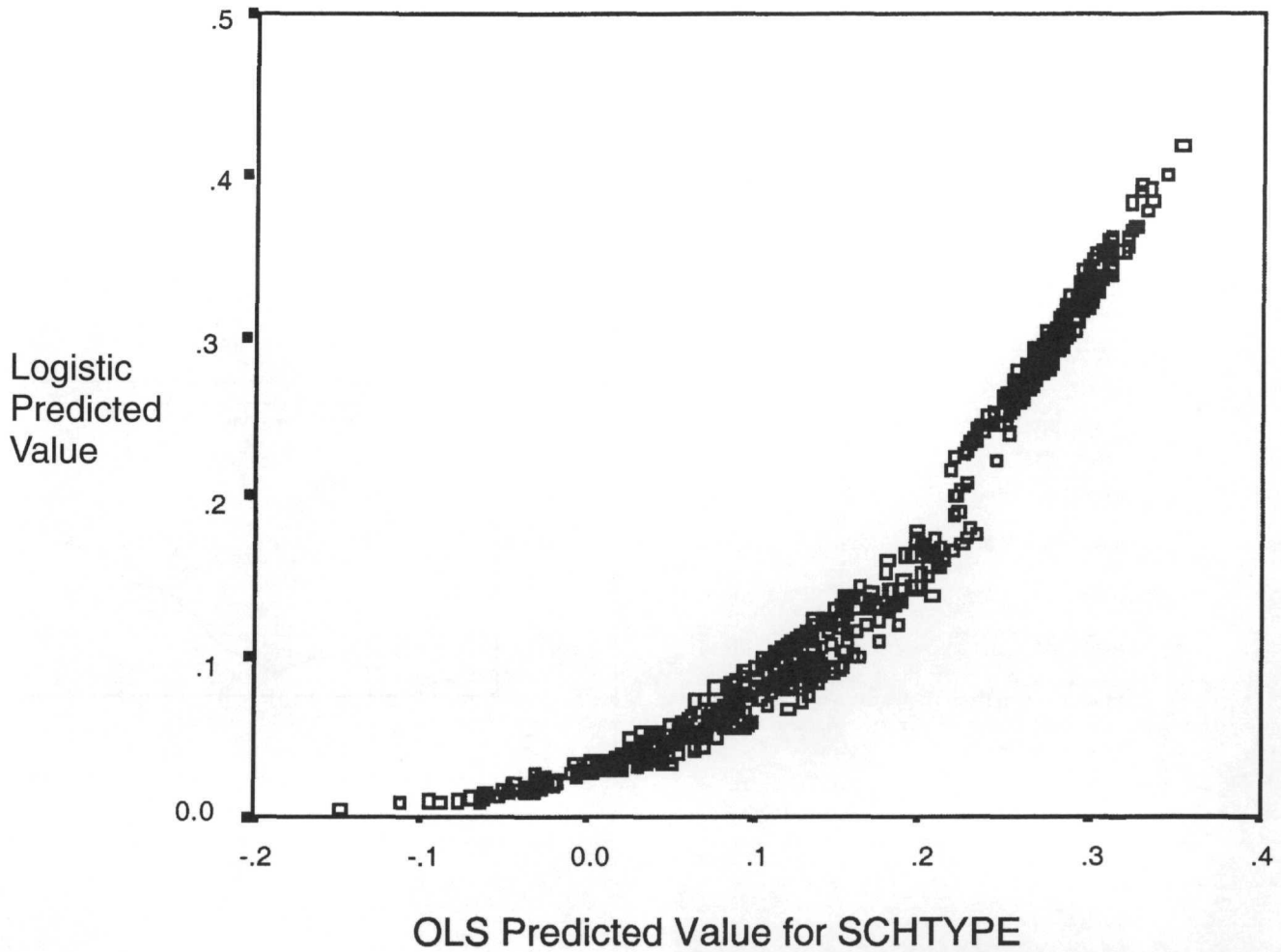


FIGURE 2. Scatter plot of predicted values for OLS and logistic regression for the high school and beyond data (n = 600).

examined here. If a researcher were concerned only with testing relationships, either model could be used. Both OLS and logistic regression yielded the same significance test results (alpha = 0.05) for the entire model and each independent variable. One would make the same substantive interpretations using either model.

In both data sets, logistic regression produced more accurate estimates of the probability of belonging to the dependent category. Negative OLS predicted values were observed on a number of cases in both studies, and the OLS predicted values were not as strongly related to the dependent binary variable as were the logistic estimates. Also, the logistic estimates were aligned more closely with observed probabilities compared to the OLS estimates. If the purpose of the research is estimating probabilities of the outcome event, logistic regression is the better model.

While these results are based on only two data sets, the pattern of findings were identical. These findings could be discerned from a careful analysis of the structural models underlying logistic and OLS regression. The OLS prediction of Y and the logistic prediction of the log odds that Y = 1 are monotonically related. The logistic function is restrained to range between 0 and 1, while OLS predictions are not so constrained. Nonlinear transformations of the independent variables in OLS, such as polynomial expansions, could be used to fit nonlinear relationships. But, OLS could still give predicted values outside the (0,1) range. There is no way to limit the OLS predicted values to this range and still satisfy unconditionally the least squares criterion.

If the purpose of the analysis was to classify cases on the dependent variable outcome, either model could be used. Both models yielded almost identical classifications of students as dropouts or private college attendees.

A review of all the results suggests that logistic regression should be used to model binary dependent variables. The structure of the logistic regression model is designed for binary outcomes, whereas OLS is not. Logistic regression results will be comparable to those of OLS in many respects, but give more accurate predictions of probabilities on the dependent outcome. Social science researchers should become more familiar with logistic regression methods and begin to use them when modeling the probability of binary outcomes.

LITERATURE CITED

Cox DR, Snell EJ. 1989. The Analysis of Binary Data. 2nd ed. London: Chapman and Hall. 236 p.

Hocking RR. 1985. The Analysis of Linear Models. Monterey: Brooks/Cole. 385 p.

Hosmer DW, Lemeshow S. 1989. Applied Logistic Regression. New York: John Wiley & Sons. 307 p.

Howell DC. 2002. Statistical Methods for Psychology. 5th ed. Pacific Grove (CA): Duxbury. p 728-31.
King EN, Ryan TP. 2002. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. American Statistician 56(3):163-70.
Menard S. 1995. Applied logistic regression analysis. Sage Univ Paper series on Quantitative Applications in the Social Sciences series no. 07-106. Thousand Oaks (CA): Sage. 98 p.
National Center for Education Statistics. 1995. Statistical Analysis Report January 1995 High School and Beyond: 1992 Descriptive Summary of 1980 High School Sophomores 12 Years Later. Washington: US Dept of Education. 112 p.
SAS Institute Inc. 1989. SAS/STAT User's Guide. Version 6, 4th ed., vol. 2. Cary (NC): SAS Institute. 846 p.