Estimating Models with Binary Dependent Variables: Some Theoretical and Empirical Observations

Guy Gessner
Canisius College

Wagner A. Kamakura
Vanderbilt University

Naresh K. Malhotra
Georgia Institute of Technology

Mark E. Zmijewski
University of Chicago

Many mathematically similar models are being used by business researchers to link binary dependent variables with a set of predictor variables. Typical research results indicate little difference between models in their ability to properly classify observations. But, there appear to be major differences in the interpretation of coefficients resulting from the calibration of these competing models.

The empirical results in this article clearly show that when the assumptions underlying binary-dependent-variable techniques are violated, parameter estimates may be misleading. This can be true even when the goodness-of-fit statistics are not substantially affected.

Introduction

Business researchers in several areas are concerned with estimating models with binary dependent variables. Binary dependent variables arise most often in classification problems. The typical classification problem in business requires the researcher to predict, from a set of independent variables, in which category a stimuli belongs. These nominal, dependent categories could be competing brands of a product or service [30], competing modes of travel [5], creditworthiness [33], labor-force participation [13], or financial solvency [14, 40].

The need to predict category or group membership is a common problem to almost all types of business researchers. In this context, many researchers have tried a wide variety of estimation methods to link independent variables with nominal dependent variables. Some researchers have found that a range of alternative techniques produce similar abilities to classify observations correctly [13,

Address correspondence to Professor Naresh K. Malhotra, College of Management, Georgia Institute of Technology, Atlanta, Georgia 30332.

Journal of Business Research 16(1), 49–65 (1988)
© 1988 Elsevier Science Publishing Co., Inc. 1988
52 Vanderbilt Ave., New York, NY 10017
0148-2963/88/$3.50

35, 38]. Others have favored a logit approach [22, 23, 28, 29]. Finally, other researchers [11, 37] have found the probit model to be superior.

Although the classification problem has been thoroughly researched and reported, the properties and interpretation of parameter coefficients has not. There appears to be an ongoing disagreement as to the importance of the effects of technique-assumption violations on parameter estimation [34, 17, 9]. However, a preponderance of empirical and theoretical evidence has been compiled by econometricians [18, 23], market researchers [19], and psychometricians [36, 39] that indicates that researchers need to pay close attention to violations in the assumptions underlying statistical models.

Ingram and Frazier [14] highlight the extent of coefficient variability between these alternative estimation techniques: "The results indicate that while there were only small differences in the classifying accuracy of the three approaches, substantially different (even opposite) conclusions were supported regarding the significance of the individual variables." If these conclusions are correct, then the choice of estimation model is crucial for hypothesis testing and statistical inference.

The problem of coefficient variability in linear regression has long been a concern of econometricians. Econometricians have identified several data-specific problems that affect coefficient interpretation and have proposed a set of numeric solutions to these problems [2]. Each of the data-specific problems commonly encountered in econometric problems results in a violation of the underlying assumptions of the linear- regression model. In these problem situations, the regression coefficients are no longer best, linear, unbiased estimators (BLUE).

Zmijewski [40] has examined the estimation biases that result from nonrandom samples. This article is an extention of this examination. The effects are explored of certain commonly encountered assumption violations on five popular techniques used with binary dependent variables. Examined in this article are the effects of three assumption violations:

1. Unequal group variance–covariance matrices
2. Nonnormal joint distribution of the independent variables
3. Multicollinearity of the independent variables

on five estimation techniques:

1. Linear Discriminant Analysis (LDA)
2. Binary Logit Analysis
3. Ordinary Least Squares (OLS)
4. Binary Probit Analysis
5. Quadratic Discriminant Analysis (QDA)

A simulation approach is explored to control and clearly demonstrate the effects of assumption violations on each technique. Using actual corporate bankruptcy data, the prevalence of these assumption violations are shown, and their effects on the interpretation of empirical results are described. The results indicate that the simultaneous existence of both nonnormally distributed and highly collinear predictor variables can result in estimation differences across these techniques. Thus, researchers should attempt to transform data so that these characteristics are not present when estimating such models.

Conceptual Discussion

The choice of the appropriate technique to estimate models with binary dependent variables depends upon the characteristics of the data. Each technique has a set of underlying assumptions that must *not* be violated by the data for the appropriate estimation and testing of the model parameters. A summary of the major underlying assumptions of the five techniques appears in Table 1. If the assumptions in Table 1 are not violated, all five techniques will provide qualitatively equivalent estimates of the model [18]. However, violations of one or more of these assumptions can result in significant differences across techniques.

Fisher [6] demonstrated that a direct analogy exists between two-group LDA and the special case (linear-probability model) of OLS: The regression estimates are proportional to the LDA "weights" by a factor of SSE/(n_1 + n_2 + 2); where SSE is the sum of the squared residuals and n_1, n_2 are the sample sizes of each group. However, due to the problem of heteroscedasticity in the regression residuals, the OLS estimates are inefficient unless a generalized least-squares approach is used [18]. The equivalence of LDA, logit analysis, and OLS has been demonstrated under specific conditions [24].

OLS estimates should be approximately one-fourth as large (in absolute value) as the logit coefficients (with the exception of the intercept) [1]. Similar empirical equivalences can also be obtained for the binary probit model. The major difference between the logit and probit models is the assumption regarding the error distribution: the logit model assumes a skewed Weibull distribution as compared to probit's symmetric normal-distribution assumption. The parameters from the logit model can be compared to the probit estimates after a correction is made for the difference in the variances between the two distributions. Since the logistic distribution has a fixed variance of $3/\pi^2$, and the probit model uses a standard normal distribution, the logit estimates multiplied by a factor of $\sqrt{3}/\pi$ are comparable to the probit estimates [1]. The logit and probit forms are similar with minor differences being found in the tails of these distributions [19].

The discussion indicates that these techniques should provide qualitatively equivalent empirical results. However, authors [14] who compared LDA, logit, and probit empirically, did not find equivalent results. In fact, many of the coefficients reported have statistically significant positive estimates using one technique and statistically significant negative coefficients using another technique. The most likely explanation for these results is that some of the underlying assumptions are violated for at least one of the techniques they examine.

LDA assumes 1) the predictor variables are distributed multivariate normal and 2) the distribution of predictors across groups differ only by a shift in means, that is, they have identical variance–covariance matrices. If these conditions are satisfied, and if the population frequency rates for both groups are known, then LDA provides the optimal classification rule and the discriminant weights are the true maximum-likelihood estimates (MLE) of the discriminant function. Otherwise, these estimates are neither efficient nor consistent. On the other hand, the MLE that are obtained either through the binary-logit or probit models are consistent even if these conditions do not prevail [18]. The robustness of LDA has received the attention of many researchers (see [16] for a comprehensive review).

When the covariance matrices of the distribution of predictors for the two groups

Table 1. Summary of Underlying Assumptions of Five Statistical Techniques

	Statistical Techniques				
Assumption	Linear Discriminant	Logit Analysis	OLS Regression	Probit Analysis	Quadratic Discriminant
1. Multivariate distribution of predictor variables	Normal	N/A[a]	N/A	N/A	Normal
2. Variance–covariance Matrix of predictor Variables across groups	Equal	Both Diagonal	Both Diagonal	Both Diagonal	Unequal
3. Distribution of error term	N/A	Weibull	Normal	Normal	N/A

[a] N/A: Assumption does not apply.

are unequal, QDA provides the optimal classification rule. A special case was examined [10] in which the covariance matrix for one group is the identity matrix ($\Sigma_1 = I$), and the second group has a covariance matrix proportional to the first ($\Sigma_2 = d\mathrm{I}$). Several conditions were considered [10] by varying the distance between groups, the relative size of the groups, and the proportionality factor, *d*. The results indicate that QDA provides superior results as *d* increases and as the difference between groups decreases. Marks and Dunn [21] showed that QDA results are substantially poorer than LDA's for small samples.

Violations of the normality assumption have been studied at two levels; discrete distributions and continuous but nonnormal distributions. Some researchers [16] show that LDA performs reasonably well for predictors with a discrete distribution. Regarding continuous nonnormal distributions, fewer studies appear in the literature. Lachenbruch, Sreeringer, and Revo [16] considered three distributions obtained from transformations of normal variates. Their results showed a poor performance of both LDA and QDA in contrast with the reasonable performance of these models with discrete distributions. Although there is no multivariate distributional assumption for the predictor underlying logit and probit analysis, strict error-term distributional assumptions underlie these techniques.

Another important issue in the specification of a discriminant function is the problem of multicollinearity among predictors (i.e., nondiagonal covariance matrices). Multicollinearity produces effects in the discriminant-analysis context that might seem counterintuitive to the econometrician [3]. Let $D^2{}_2$ be the distances between two groups on predictors x_1 and x_2. The collinearity between x_1 and x_2 will improve discrimination (as compared with the independent case) if,

$$(D_2 - rD_1)^2/(1-r^2)>D_2^2 \tag{1}$$

Hence, the fact that two predictors are correlated might improve discrimination. In fact, a negative correlation between x_1 and x_2 will always improve discrimination over the independent case [12]. Logit, OLS, and probit analyses, on the other hand, are all affected by collinearity in a similar manner. That is, the parameter estimates are unbiased but inefficient, and the statistical significance of individual coefficients cannot be determined separately.

The relative advantage of one statistical technique over others seems to depend on the conditions (i.e., characteristics of the data) under which these methods are applied, and to the extent that these conditions violate the assumptions underlying each technique. This issue is examined empirically; first by using simulated data and then by using actual data.

Empirical Comparisons Based on Simulated Data

Two sets of simulations are conducted. Both simulations estimate the same underlying model.

$$y_i = F(x_{1i}, x_{2i}) \tag{2}$$

where

y_i = 1 if observation i is a member of group A and 0 if observation i is a member of group B,
$F(\cdot)$ = the functional relationship between y and x_1 and x_2,
x_{1i} = predictor variable 1 for observation i, and
x_{2i} = predictor variable 2 for observation i.

The simulations differ primarily in the multivariate distribution of the predictor variables. In the first set of simulations, the predictor variables are bivariate normals. The predictor variables are distributed log-normally in the second set of simulations.

In the first set of simulations (the bivariate normal simulations), group A predictors (x_1, x_2) have a *standard* bivariate normal distribution (i.e., both predictors have means equal to 0.0 and variances equal to 1.0) and a correlation coefficient ($r_{x_1x_2}(A)$) of either 0.0 or 0.9, depending on the combination of attributes that are examined. Group B predictors have means of $\bar{x}_1(B) = 1.0$ and $\bar{x}_2(B) = .05$, variances of $S^2_{x_1}$ $(B) = 1.0$ or 4.0 and $S^2_{x_2}$ $(B) = 1.0$, and a correlation coefficient ($r_{x_1x_2}$ (B)) of 0.0 or 0.9, depending on the combination of attributes that are examined. By varying $S^2_{x_1}(B)$, $r_{x_1x_2}(A)$, and $r_{x_1x_2}(B)$, we examine the effect of unequal covariances and multicollinearity on these techniques to demonstrate potential empirical differences.

An estimation sample and a prediction sample containing 600 observations (300 observations in each group), are used to compare the five statistical techniques across six combinations of the sample attributes: ($r_{x_1x_2}(A) = 0.0, 0.9$; $S^2_{x_2}$ $(B) = 1.0, 4.0$; $r_{x_1x_2}(B) = 0.0, 0.9$). Each of these samples was randomly generated and are denoted as simulations 1.1 through 1.6. The results of these simulations are presented in Table 2. Panel A of Table 2 describes the combination of sample attributes for each of the simulations. Simulation 1.1 is the case where the data are multivariate normal, the groups have equal variances and covariances, and the predictors are uncorrelated (i.e., no assumptions are violated). Simulation 1.2 is the case where the group covariances are unequal. Simulations 1.3 and 1.4 do not violate the multivariate normality or equal assumptions, but the predictor variables are correlated in one (simulation 1.3) or both (simulation 1.4) groups. Finally, simulations 1.5 and 1.6 violate the covariance matrix-equality assumption and have correlated predictor variables in one (simulation 1.5) or both (simulation 1.6) groups.

Panel B of Table 2 includes the estimated coefficients for LDA, Logit, OLS, and Probit. While there exist differences in the parameter estimates, these differences are consistent with the estimation differences we discussed in Section 2. The last row in Panel B shows the ratio of the coefficient for x_1 to the coefficient for x_2. Since the parameter estimates across techniques are comparable using linear transformations, the ratios of the parameters can be used for comparing techniques. These results indicate that there is no substantial difference in these ratios across statistical techniques and hence, the parameter estimates are as expected from the conceptual discussion above.

Panels C and D of Table 2 report the percentage of observations that are classified correctly for the estimation sample and predicted correctly for the prediction sample. Qualitatively, Panels C and D provide equivalent results, hence, only the classification results are discussed. Comparing simulations 1.1 to 1.2 indicates that

Table 2. LDA, Logit, OLS, Probit, and QDA Comparisons: Simulated, Normally Distributed Data

Panel A: Sample Attributes	Simulation					
	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**
s^2 [*B*]	1.0	4.0	1.0	1.0	4.0	4.0
*r*x1 [*A*]	0.0	0.0	0.9	0.9	0.9	0.9
*r*x1x2 [*B*]	0.0	0.0	0.0	0.9	0.0	0.9
*x*1*x*2						
			Panel B: Estimated Parameters			
Intercept:						
LDA	–0.59	–0.33	–0.52	–0.69	–0.31	–0.28
Logit	–0.70[a]	–0.21[a]	–0.56	–0.96[a]	–0.19[a]	–0.16[a]
OLS	0.35	0.44	0.38	0.31	0.45	0.47
Probit	–0.41[a]	–0.12[a]	–0.33[a]	–0.57[a]	–0.12[a]	–0.10[a]
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Coefficient for *x*1:						
LDA	0.90	0.21	0.95	2.13	0.18	–0.06
Logit	1.06[a]	0.14[a]	1.00[a]	2.95[a]	0.11[a]	–0.04
OLS	0.22[a]	0.03[a]	0.21[a]	0.56[a]	0.03[a]	–0.01
Probit	0.63[a]	0.08[a]	0.60[a]	1.75[a]	0.07[a]	–0.02
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Coefficient for *x*2:						
LDA	0.48	0.82	0.10	–1.57	0.82	1.15
Logit	0.57[a]	0.53[a]	0.12	–2.17[a]	0.50[a]	0.63[a]
OLS	0.12[a]	0.12[a]	0.03	–0.41[a]	0.12[a]	0.15[a]
Probit	0.34[a]	0.33[a]	0.07	–1.29[a]	0.31[a]	0.39[a]
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
b[*x*1]/*b*[*x*2]:						
LDA2	1.87	0.25	9.60	–1.36	0.22	–0.05
Logit	1.86	0.26	8.65	–1.36	0.22	–0.06
OLS	1.84	0.27	8.00	–1.37	0.22	–0.05
Probit	1.88	0.26	9.20	–1.36	0.22	–0.05
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
			Panel C: Classifications			
Group A:						
LDA	73%	64%	70%	74%	62%	59%
Logit	72%	64%	70%	74%	62%	59%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	72%	63%	70%	74%	62%	60%
QDA	73%	76%	69%	74%	69%	52%
Group B:						
LDA	71%	62%	69%	77%	62%	61%
Logit	71%	62%	69%	77%	62%	61%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	71%	62%	69%	77%	62%	61%
QDA	70%	52%	70%	77%	56%	66%
Total:						
LDA	72%	63%	70%	75%	62%	60%
Logit	72%	63%	70%	75%	62%	60%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	72%	63%	70%	75%	62%	60%
QDA	72%	64%	69%	75%	62%	59%

Table 2. *continued*

	Panel D: Predictions					
Group A:						
LDA	76%	64%	72%	73%	65%	63%
Logit	76%	64%	72%	74%	65%	63%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	76%	63%	72%	73%	65%	63%
QDA	76%	79%	72%	73%	73%	58%
Group B:						
LDA	73%	61%	70%	77%	60%	59%
Logit	73%	62%	70%	76%	60%	59%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	74%	63%	70%	77%	61%	59%
QDA	72%	47%	70%	77%	54%	67%
Total:						
LDA	75%	63%	71%	75%	63%	61%
Logit	75%	63%	71%	75%	62%	61%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	75%	63%	71%	75%	63%	61%
QDA	74%	63%	71%	75%	64%	62%

[a]Significant at the 95% confidence interval.
Significant tests for LDA are not reported.

$S_{x_1}^2(J)$ = the variance of predictor $x(1)$ in group J (J = A, B), all other variances are fixed.
$r_{x_1x_2}(J)$ = the correlation coefficient between $x(1)$ and $x(2)$ in group J (J = A, B).
LDA = Linear Discriminant Analysis.
QDA = Quadratic Discriminant Analysis.
OLS = Ordinary Least Squares. Criteria coded (0 or 1) based upon group membership.
n.a. = not applicable.

unequal covariances diminishes the predictive ability of all techniques. Comparsimulations 1.1 to 1.4 and 1.2 to 1.6 indicates that collinearity among the preor variables can either increase or decrease predictive ability, depending on the direction of the inequality that appears in Equation (1). When $(D_2-rD_1)^2/(1-r^2)>D_2$, then predictive ability is increased; if the inequality is reversed, then predictive ability is decreased. For example, in simulation 1.4 $(D_2-rD_1)^2/(1-r^2)$ = .84 and D_2^2 = .25, thus, we would expect collinearity to improve discrimination. In simulation 1.6, on the other hand, $(D_2-rD_1)^2/(1-r^2)$ = 0.13 and D_2^2= .25, thus, we expect collinearity to reduce the discriminatory power. This effect is demonstrated in Table 2 for both the classifications and predictions. However, note that the data characteristics (e.g., predictor-variable correlations) in the classification and prediction samples are always identical. This may not be the case when using actual data and hence, similar results may not be observed in both classification and prediction samples.

The important conclusion from these analyses is that the five statistical techniques provide equivalent results empirically if: 1) the data do not violate any of the underlying assumptions, 2) the group covariance matrices are unequal, and/or 3) the predictor variables are collinear. Thus, the simulation tests so far are not able to explain the empirical differences previously reported across these techniques [14].

The second set of simulations, 2.1 through 2.6, are analogous to the fist set of simulations (1.1 through 1.6) with one exception; the predictor variables are trans-

Table 3. LDA, Logit, OLS, Probit, and QDA Comparisons: Simulated, Log-Normally Distributed Data

Panel A: Sample Attributes	Simulation					
	2.1	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**
s^2 [*B*]	1.0	1.5	1.0	1.0	1.5	1.5
*rx*1 [*A*]	0.0	0.0	0.9	0.9	0.9	0.9
*rx*1*x*2 [*B*]	0.0	0.0	0.0	0.9	0.0	0.9
*x*1*x*2						
			Panel B: Estimated Parameters			
Intercept:						
LDA	–1.04	–0.83	–0.98	–0.48	–0.83	–0.41
Logit	–1.41*	–1.25*	–1.17*	–0.89*	–1.06*	–0.41*
OLS	0.22	0.30	0.23	0.34	0.31	0.42
Probit	–0.81*	–0.74*	–0.64*	–0.51*	–0.59*	–0.24*
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Coefficient for *x*1:						
LDA	0.18	0.05	0.18	0.37	0.05	0.06
Logit	0.40*	0.30*	0.37*	1.13*	0.27*	0.82*
OLS	0.07*	0.02*	0.07*	0.16*	0.02*	0.03*
Probit	0.22*	0.17*	0.19*	0.59*	0.15*	0.48
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Coefficient for *x*2:						
LDA	0.22	0.27	0.20	–0.32	0.28	0.04
Logit	0.20*	0.20*	0.13*	–0.97*	0.15*	–0.95*
OLS	0.05*	0.05*	0.04*	–0.13*	0.05*	–0.02
Probit	0.12*	0.12*	0.05*	–0.51*	0.07*	–0.57*
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
b[*x*1]/*b*[*x*2]:						
LDA	0.85	0.19	0.90	–1.15	0.16	1.28
Logit	1.98	1.46	2.94	–1.17	1.83	–0.86
OLS	1.47	0.46	1.67	–1.20	0.44	–1.38
Probit	1.81	1.44	3.88	–1.16	2.27	–0.84
QDA	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
			Panel C: Classifications			
Group A:						
LDA	86%	85%	85%	92%	83%	94%
Logit	82%	84%	82%	84%	85%	88%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	83%	85%	83%	85%	86%	88%
QDA	93%	92%	91%	94%	92%	97%
Group B:						
LDA	53%	46%	52%	46%	46%	36%
Logit	60%	54%	57%	62%	54%	52%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	57%	54%	53%	58%	52%	52%
QDA	38%	28%	35%	38%	28%	30%
Total:						
LDA	69%	65%	69%	69%	65%	65%
Logit	71%	69%	69%	73%	70%	70%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	70%	70%	68%	72%	69%	70%
QDA	66%	60%	63%	66%	60%	64%

Table 3. *continued*

	Panel D: Predictions					
Group A:						
LDA	84%	84%	84%	95%	84%	92%
Logit	81%	83%	82%	85%	83%	91%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	81%	83%	83%	88%	85%	91%
QDA	91%	92%	90%	94%	91%	95%
Group B:						
LDA	48%	45%	48%	44%	45%	31%
Logit	57%	53%	57%	60%	53%	51%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	56%	53%	54%	58%	50%	50%
QDA	39%	32%	36%	37%	32%	27%
Total:						
LDA	66%	65%	66%	69%	65%	62%
Logit	69%	68%	70%	73%	68%	71%
OLS	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Probit	68%	68%	69%	73%	68%	71%
QDA	65%	62%	63%	66%	62%	61%

formed so that they are log-normally distributed. The results of these simulations are reported in Table 3 and indicate some differences across estimation techniques. First, the simulation with unequal covariance matrices and collinear data (simulation 2.6) has coefficients (Panel B of Table 3) of different signs for predictor variable x_2. The ratios of the coefficients (see the last set of rows in Panel B) also indicate differences across estimation techniques. Second, the discriminant models (LDA and QDA) tend to have less predictive ability, especially for group *B*.

These results provide a possible explanation for previous results [14] and identify a set of conditions of which researchers should be aware. The techniques researchers commonly use to estimate models with binary dependent variables can result in different statistical inferences and classifications if the predictor variables are not normally distributed and the data are highly collinear. Thus, researchers should be aware of these problems when estimating such models since many predictor variables used in research are not normally distributed and are often collinear (e.g., security prices, security returns, and financial ratios). Actual data are examined next that are not normally distributed and are highly collinear to determine if the same results are observed using actual data and also to determine if eliminating the collinearity can ameliorate this problem.

Empirical Comparisons Based on Actual Data

The empirical comparison reported is based upon a bankruptcy-prediction model similar to the numerous such models appearing in the literature. The empirical investigation was designed to complement the simulation studies. The simulations examined only two predictor variables. This empirical study includes a greater number of predictors. Two alternative models are used as a basis for the tests; one model consists of nine variables (some of which are highly correlated) and the other consists of three orthogonalized factors. A comparison between these two models provides some insights as to the effects of multicollinearity on these techniques using financial ratios as predictor variables (which we know are not normally distributed [7]). The sample used in this analysis came from the following sources: For bankrupt firms, 10K reports were available over the 1972 to 1978 time period under study including the last report filed with SEC before the filing of the bankruptcy petition (found in the Cornell University 10K collection). A total of 129 firms have filed petitions for bankruptcy. Complete data are available for 72 of the firms. The sample of nonbankrupt firms consisted of all firms that are contained on the 1979 COMPUSTAT Annual Industrial Tape and have complete data over this period. A total of 3,573 firms were included in the nonbankrupt sample.

The first model contains three *sets* of highly correlated variables, each attempting to measure a particular financial characteristic: liquidity, financial leverage, and rate of return. (See Table 4 for a description of these variables). Descriptive statistics and a correlation matrix for these variables are reported in Table 4. An examination of the groups' means (see Panel A, Table 4) indicates that the bankrupt firm sample has, on average, lower liquidity measures, higher levels of financial leverage, and lower rates of return. The correlations matrix (see Panel B, Table 4) indicates that the variables are highly correlated within the three financial characteristic measures. Thus, the construction of this model induces problems of severe multicollinearity.

The results from estimation for the five statistical techniques are reported in Table 5. All five estimations indicate that the overall model is statistically significant. An examination of the individual coefficients indicates the potential existence of multicollinearity, (i.e., the various measures of the financial characteristics have different signs with some (but not all) coefficients significantly different from zero). An *F* ratio testing the null hypothesis of group variance–covariance matrix equality is 68.905 with 45 and 56,113 degrees of freedom. This value rejects the null hypothesis at less than the .001 level. Thus, the LDA assumptions are violated and the QDA is more appropriate.

An examination of the coefficients across methodologies indicates similar results to those reported previously [14]. Both the signs of the coefficients and the statistical inferences across methodologies are inconsistent. Examing the percentage of firms classified correctly across methodologies indicates that, overall, the techniques appear to classify firms equally well. However, there are some differences within individual groups. LDA appears better at classifying nonbankrupt firms, but it does not have an overall correct classification rate significantly different from the other techniques.

In an attempt to eliminate the potential problem of multicollinearity, another model was tested that consists of three orthogonalized factors. The nine variables from the above model are factor analyzed and factor scores are computed for each observation. As one would expect, the three sets of variables (liquidity, financial leverage, and rate of return) "loaded" almost uniquely on the first three factors. The first three factors explain 93% of the total variance, with the rate-of-return factor explaining 44%, the liquidity factor 33%, and the financial-leverage factor 16%. The remaining six factors have little ability to explain any of the remaining variance. These first three factors, which are orthogonal by design, are used to formulate the factor-score (orthogonal) model.

The estimation results using the alternative methodologies for the factor-score

Table 4. Variables for Comparing Statistical Techniques on Actual Data: Descriptive Statistics and Correlation Coefficients

Panel A: Descriptive Analysis

	Bankrupt		Nonbankrupt		Overall	
Variables	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Liquidity						
1. CA to CL[a]	1.87	21.97	2.84	24.42	2.82	24.13
2. CASH to CL	.12	20.34	.84	23.98	.82	23.69
3. (CA–INV) to CL	.99	6.45	1.64	18.41	1.63	18.19
Financial Leverage						
4. ID to TA	.90	.67	.48	.16	.49	.20
5. (TD+PD) to TA	.90	.67	.49	.17	.50	.21
6. LTD to TA	.33	.25	.21	.14	.23	.14
Rate of Return						
7. NI to TA	–.16	.29	.05	.07	.05	.09
8. CF to TA	–.10	.35	.16	.06	.15	.12
9. EBIT to TA	–.13	.36	.12	.10	.12	.12

Panel B: Correlations

Variables	1	2	3	4	5	6	7	8	9
Liquidity									
1. CA to CL[a]	1.00	1.00	1.00	–.06	–.06	–.02	.04	.01	.04
2. CASH to CL		1.00	1.00	–.04	–.04	–.02	.03	.01	.03
3. (CA–INV) to CL			1.00	–.06	–.06	–.02	.04	.01	.04
Financial Leverage									
4. TD to TA				1.00	–.99	.55	–.48	–.45	–.47
5. (TD+PS) to TA					1.00	.56	–.49	–.45	–.47
6. LTD to TA						1.00	–.27	–.23	–.26
Rate of Return									
7. NI to TA							1.00	.94	.96
8. CF to TA								1.00	.98
9. EBIT to TA									1.00

[a] CA = current assets.
CL = current liabilities.
CASH = cash plus marketable securities.
INV = inventory.
TD = total debt.
TA = total assets.
PS = preferred stock.
LTD = long-term debt.
NI = net income.
CF = cash flow.
EBIT = earnings before interest and taxes.

model are reported in Table 6. The overall model is statistically significant across all techniques. The F ratio testing the null hypothesis of group variance–covariance matrix equality is 237.44 with 6 and 93,968 degrees of freedom again rejecting the null hypothesis (of equality) at less than .001 level. However, contrary to previous results [14], the signs of the coefficients and the statistical inferences are consistent across all estimation techniques. The alternative methodologies appear to classify firms equally well, and LDA better classified the nonbankrupt firms, but it does not classify the total sample better than any other technique.

Table 5. LDA, Logit, OLS, Probit, and QDA Comparisons: Nine Variable Model—Actual Data

	LDA[b]	QDA[c]	Logit	Probit	OLS
Liquidity					
1. CA to CL[a]	$-.14^{R=4}$		.28	.02	−.00
2. CASH to CL	$-.14^{R=7}$		−1.89	−1.15[f]	−.00[h]
3. (CA–INV) to CL	$.38^{R=5}$		.08	.23	.01[h]
Financial Leverage					
4. ID to TA	$20.38^{R=8}$		16.80[f]	4.65[g]	.40[f]
5. (TD+PS) to TA	$-11.36^{R=8}$		−8.74[g]	−1.54	−.22[f]
6. LTD to TA	$-3.69^{R=1}$		−.84	−.43	−.07[f]
Rate of Return					
7. NI to TA	$-33.72^{R=6}$		25.99[f]	8.22	−.66[f]
8. CF to TA	$-13.57^{R=2}$		−17.88[h]	−8.34[g]	−.27[g]
9. EBIT to TA	$20.51^{R=3}$		−14.52[f]	−2.75	.40[f]
Constant	−6.43		−3.97[f]	−2.94[f]	−.02[h]
F ratio	71.65[f]	71.65[f]	—	—	71.67[f]
−2x log of Likelihood Ratio[e]	—	—	154.81[f]	318.64[f]	—
R-Squared	—	—	.26	.20	.17
Percent Correctly Classified:					
Bankrupt	48	34	38	40	—
Nonbankrupt	99	99	99	99	—
Overall	98	98	98	98	—

[a] CA = current assets.
CL = current liabilities.
CASH = cash plus marketable securities.
INV = inventory.
TD = total debt.
TA = total assets.
PS = preferred stock.
LTD = long-term debt.
NI = net income.
CF = cash flow.
EBIT = earnings before interest and taxes.

[b] R = rank as estimated via the conditional deletion procedure – see Karson and Martell (1980).

[c] QDA = quadratic equation results (54 parameters) are not reported.

[d] A small sample of nonbankrupt firms was used because the likelihood function would not converge with the entire data set.

[e] Distributed as a chi-square with 9 degrees of freedom.

[f] Significant at less than the .01 level.

[g] Significant at less than the .05 level.

[h] Significant at less than the .10 level.

Summary and Conclusions

It is clear from the results that the choice of statistical technique should be data dependent. Table 7 and Table 8 present a summary of data conditions and technique performance. Statisticians have often reminded social scientists of the need for identification of the properties of the data prior to selection of an estimation technique [4, 26, 31].

Each technique numerically calibrates a model that is based upon a set of assumptions about the data. Violations of these underlying assumptions cause esti-

Table 6. LDA, Logit, OLS, Probit, and QDA Comparisons: Three Orthogonal Factor Model—Actual Data

	LDA[a]	QDA[b]	Logit	Probit	OLS
1. Liquidity	$-.03^{R=3}$	R=3	.02	.01	.00
2. Rate of return	$-2.28^{R=1}$	R=1	-1.15[c]	-.49[c]	-.05[c]
3. Financial leverage	$1.91^{R=2}$	R=2	1.27[c]	.57[c]	.04[c]
Constant	-3.55		-4.80[c]	-2.44[c]	-.02[c]
F ratio	186.36	186.36	—	—	186.37
-2x Log of Likelihood Ratio[d]	—	—	264.88	266.84	—
R-squared	—	—	.26	.21	.15
Percent Correctly Classified:					
Bankrupt	48	34	48	38	—
Nonbankrupt	99	99	99	99	—
Overall	98	98	98	98	—

[a] R = rank as estimated via the conditional deletion procedure - see Paksoy et. al. (1977).
[b] QDA = quadratic equation results (9 parameters) are not reported.
[c] Significant at the .01 level.
[d] Distributed as a chi-square with 3 degrees of freedom.

Table 7. Data Conditions and Technique Performance

	Predictor Characteristics					
Criteria	Distribution	Means	Variances	Correlations	Group Covariances	Data Condition
Group 1	Normal	Equal	Equal	Orthogonal	Equal	1
Group 2	Normal	Unequal	Equal	Orthogonal		
Group 1	Normal	Equal	Equal	Orthogonal	Equal	2
Group 2	Normal	Unequal	Unequal	Orthogonal		
Group 1	Normal	Equal	Equal	Collinear	Unequal	3
Group 2	Normal	Unequal	Equal	Orthogonal		
Group 1	Normal	Equal	Equal	Collinear	Equal	4
Group 2	Normal	Unequal	Equal	Collinear		
Group 1	Normal	Equal	Equal	Collinear	Unequal	5
Group 2	Normal	Unequal	Unequal	Orthogonal		
Group 1	Normal	Equal	Equal	Collinear	Equal	6
Group 2	Normal	Unequal	Unequal	Collinear		

			Techniques That Performed Best:	
Data Condition	Coefficient Size	Ratio Signs	Correct Classifications	Correct Predictions
1	All same	All (+)	All equal	LDA, Logit, Probit
2	All same	All (+)	QDA	All equal
3	All same	All (+)	All equal	All equal
4	All same	All (−)	All equal	All equal
5	All same	All (+)	All equal	QDA
6	All same	All (−)	LDA, Logit, Probit	QDA

Table 8. Data Conditions and Estimation Technique Performance

	Predictor Characteristics					
Criteria	Distribution	Means	Variances	Correlations	Group Covariance	Data Condition
Group 1	Log-normal	Equal	Equal	Orthogonal	Equal	1
Group 2	Log-normal	Unequal	Equal	Orthogonal		
Group 1	Log-normal	Equal	Equal	Orthogonal	Equal	2
Group 2	Log-normal	Unequal	Unequal	Orthogonal		
Group 1	Logt-normal	Equal	Equal	Collinear	Unequal	3
Group 2	Log-normal	Unequal	Equal	Orthogonal		
Group 1	Log-normal	Equal	Equal	Collinear	Equal	4
Group 2	Log-normal	Unequal	Equal	Collinear		
Group 1	Log-normal	Equal	Equal	Collinear	Unequal	5
Group 2	Log-normal	Unequal	Unequal	Orthogonal		
Group 1	Log-normal	Equal	Equal	Collinear	Equal	6
Group 2	Log-normal	Unequal	Unequal	Collinear		

			Techniques That Performed Best On:	
Data Condition	Coefficient Size	Ratio Signs	Correct Classifications	Correct Predictions
1	Varies	All (+)	Logit & Probit	Logit, Probit
2	Varies	All (+)	Logit & Probit	Logit, Probit
3	Varies	All (+)	Logit	Logit, Probit
4	Varies	All (−)	Logit	Logit, Probit
5	Varies	All (+)	Logit	Logit, Probit
6	Varies	Varies	Logit & Probit	Logit, Probit

mation problems for each technique and an improper estimation of model coefficients.

Empirical problems (such as coefficient interpretations changing as different techniques are applied) were simulated by inducing multicollinearity and using nonnormal predictors simultaneously. Predictive ability was shown to vary as a result of the degree and direction of collinearity among predictor variables. Thus, the simulations that are performed in this paper demonstrate how each technique will vary with violations of different assumptions. Applying the knowledge gained from the simulations to the problem of predicting corporate bankruptcies, a linear transformation of collinear predictor variables to orthogonal factors eliminated the coefficient inconsistencies.

The results indicate that various techniques for analyzing binary data are likely to differ in their performance under the following conditions: the distribution of the predictor variables deviates substantially from normality, there is multicollinearity between the predictor variables, and the number of predictors is large. While these conditions are by no means exhaustive, the existence of any of these conditions should caution the researchers that the choice of a particular technique should be made carefully.

Consistent with past studies, the performance of logit and probit was similar under the various conditions. Hence, a choice between these two may not be consequential (except in computational cost). In cases where there are a large number of predictors, the QDA may not be practical. However, the choice between logit or probit, LDA and OLS is still not straightforward. Hence, the researcher should first conduct some preliminary data analysis to determine the statistical properties of the predictor variables. Perhaps part of the data could be analyzed by these techniques to determine which one is most appropriate. Alternatively, the researcher could transform the data to comply with the assumptions of a particular technique.

References

1. Amemiya, T., Qualitative Response Models: A Survey, *Journal of Economic Literature* 19(4)(1981): 483-536.
2. Belsely, D. A., Kuh, E., and Welsch, R. E., *Regression Diagnostics.* Wiley, New York, 1980.
3. Cochran, W. G., On the Performance of the Linear Discriminant Functions, *Technometrics* 6 (1964): 173–90.
4. Daniel, C., and Wood, F. S., *Fitting Equations to Data.* Wiley, New York, 1980.
5. Domencich, T., and McFadden, D., *Urban Travel Demand: A Behavioral Analysis.* North Holland, Amsterdam, 1975.
6. Fisher, R. A., The Use of Multiple Measurement in Taxonomic Problems, *Ann. Eugen.* 7 (1936): 179–88.
7. Foster, G., *Financial Statement Analysis.* Prentice-Hall, New York, 1985.
8. Frydman, H., Altman, E. I., and Kao., D., Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress, *Journal of Finance* 40 (1)(1985): 269–292.
9. Gaito, J., Measurement Scales and Statistics: Resurgency of an Old Misconception, *Psychological Bulletin* 87, no. 3, (1980): 564–567.
10. Gilbert, E. S., The Effect of Unequal Variance–Covariance Matrices on Fisher's Linear Discriminant Function, *Biometrika* 25 (1969); 505–516.
11. Grabrowsky, B. J., and Tally, W. K., Errors in the Allocation of Credit: Discriminant and Probit Analysis, *American Statistical Association 1976 Proceedings of Business and Economic Statistics Section*, 1976, pp. 326–330.
12. Green, P. E., *Analyzing Multivariate Data.* Wiley, New York, 1978.
13. Gunderson, M., Probit and Logit Estimates of Labor Force Participation, *Industrial Relations* 19, (1980): 216–220.
14. Ingram, F. J., and Frazier, E. L., Alternative Multivariate Tests in Limited Dependent Variable Models: An Empirical Assessment, *Journal of Financial and Quantitative Analysis* 27 (June 1982): 227–240.
15. Lachenbruch, P. A., *Discriminant Analysis.* Hafnew Press, New York, 1975.
16. Lachenbruch, P. A., Sreeringer, and Revo, L. T., Robusteness of the Linear and Discriminant Function to Certain Types of Non-Normality, *Common. Statistics* 1 (1975): 39–57.
17. Lord, F. M., On the Statistical Treatment of Football Numbers, *American Psychological* 8 (1953): 750–751.
18. Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, London, 1983.
19. Malhotra, N. K., A Comparison of the Predictive Validity of Procedures for Analyzing Binary Data, *Journal of Business and Economic Statistics* 1 (1983): 326–336.
20. Marais, L. M., Patell, J. M., and Wolfson, M. A., The Experimental Design of

Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank Loan Classification, Journal of Accounting Research 22 (Supplement)(1984): 82–118.

21. Marks, S., and Dunn, O. J., Discriminant Functions When Covariance Matrics are Unequal, *Journal of the American Statistical Association* 69 (1974): 555–559.
22. McCoy, J. L., and Manicke, R., A Comparison of CP Regression and Logit Analysis for Deriving Optimal Models of Mortality, *American Statistical Association 1978 Proceedings of the Social Statistics Section*, 1978, pp. 482–484.
23. McFadden, D., Conditional Logit Analysis of Qualitative Choice Behavior, in *Frontiers of Econometries*, P. Zarembka, ed., Academic, New York, pp. 105–142.
24. ———, A Comment on Discriminant Analysis Versus Logit Analysis, *Annals of Economic and Social Measurement* 5 (1976): 511–523.
25. ———, Economic Models for Probabilistic Choice Among Products, *Journal of Business* 53 (3, Part 2, July 1980): 513–529.
26. Mosteller, F., and Tukey, J. W., *Data Analysis and Regression*. Addison-Wesley, New York, 1972.
27. Paksoy, C. H., Ferguson, C. E., Karson, M., and Martell, T., MVBFS: A Program Using the Multivariate Behrens-Fisher Solution for Conditional Deletion in Two-Group Discriminant Analysis Models, *Journal of Marketing Research* 14 (May 1977): 245.
28. Powers, J. A., Marsh, L. C., Huckfelt, R. R., and Johnson, C. L., A Comparison of Logit, Probit and Discriminant Analysis in Prediction of Family Size, *American Statistical Association 1978 Proceedings of the Social Statistics Section*, 1978 pp. 693–697.
29. Press, S. J., and Wilson, S., Choosing between Logistic Regression and Discriminant Analysis, *Journal of the American Statistical Association* 73 (1978): 699–705.
30. Punj, G. N., and Staelin, R., The Choice Process for Graduate Business Schools, *Journal of Marketing Research* 15 (November 1978): 558–598.
31. Schaeffer, D. J., Olson, C., Kerster, H. W., and Janardan. K. G., The Low Dose Extrapolation Problem: A Review and a New Model, *American Journal of Mathematical and Management Sciences* 2 (3)(1982): 223–252.
32. Shapiro, S. S., and Brian, C. W., Recommended Distribution Testing Procedures, *American Journal of Mathematical and Management Science* 2 (3)(1982): 175–222.
33. Shinkel, B. A., The Effects of Limiting Information in the Granting of Credit, Ph.D. dissertation, Purdue University, 1976.
34. Stevens, S. S., On the Theory of Scales of Measurement, *Science* 103, no. 2684 (1946): 677–680.
35. Talvitie, A., Comparison of Probabilistic Mode- Choice Model: Estimation Methods and System Inputs, *Highway Research Record* 392 (1972): 111–120.
36. Thurstone, L., Psychological Analysis, *American Journal of Psychology* 38 (1980): 9–17.
37. Werner, J., Werner, W., and Budde, N., A Comparison of Probit, Logit, Discriminant and OLS: The Physicians Location Choice Problem, *American Statistical Association 1978 Proceedings of the Business and Economic Statistics Section*, 1978, pp. 631–675.
38. Wilensky, C. R., and Rossiter, L. F., OLS and Logit Study, *American Statistical Association 1978 Proceedings of the Social Statistics Section*, 1978, pp. 260–265.
39. Young, F. W., Quantitative Analysis of Qualitative Data, *Psychometrica* 46 no. 4 (1981): 357–388.
40. Zmijewski, M., Methodological Issues Related to the Estimation of Financial Distress Prediction Models, *Journal of Accounting Research* 22, supplement (1984): 59–82.