Journal of Data Science **9**(2011), 565-584

Predicting Bankruptcy with Robust Logistic Regression

Richard P. Hauser and David Booth*
Kent State University

Abstract: Using financial ratio data from 2006 and 2007, this study uses a three-fold cross validation scheme to compare the classification and prediction of bankrupt firms by robust logistic regression with the Bianco and Yohai (BY) estimator versus maximum likelihood (ML) logistic regression. With both the 2006 and 2007 data, BY robust logistic regression improves both the classification of bankrupt firms in the training set and the prediction of bankrupt firms in the testing set. In an out of sample test, the BY robust logistic regression correctly predicts bankruptcy for Lehman Brothers; however, the ML logistic regression never predicts bankruptcy for Lehman Brothers with either the 2006 or 2007 data. Our analysis indicates that if the BY robust logistic regression significantly changes the estimated regression coefficients from ML logistic regression, then the BY robust logistic regression method can significantly improve the classification and prediction of bankrupt firms. At worst, the BY robust logistic regression makes no changes in the estimated regression coefficients and has the same classification and prediction results as ML logistic regression. This is strong evidence that BY robust logistic regression should be used as a robustness check on ML logistic regression, and if a difference exists, then BY robust logistic regression should be used as the primary classifier.

Key words: Bankruptcy prediction, robust logistic regression.

1. Introduction

The prediction of corporate bankruptcy is an important and widely studied topic (Wilson and Sharda, 1994). Creditors and investors in corporations need to be able to predict the probability of default for profitable business decisions. For banks, accurate assessment of the probability of bankruptcy can lead to sounder lending practices as well as better fair value estimates of interest rates that reflect credit risks. However, the need to predict corporate bankruptcy goes beyond banks. For example, accounting firms may risk lawsuits if the auditors

*Corresponding author.

fail to issue an early warning, such as a "going concern" opinion for a troubled firm. More ubiquitous in business are derivative contracts where firms must often assess their counterparty risk. Historically much of the credit or counterparty risk assessment was to simply use ratings issued by the standard credit rating agencies. As many investors have recently discovered, these ratings tend to be reactive rather than predictive. Hence, there is a great need to develop accurate quantitative models for prediction of corporate bankruptcy.

A major approach to develop quantitative models for such prediction has been to learn the relationship of default with firm variables from data using statistical models. In both practice and in academic studies, statistical models based on multivariate discriminant analysis, logistic regression, and neural networks have been used to predict corporate bankruptcy (Sharda and Wilson, 1996; Lee *et al.*, 2005). In this study, we will focus on corporate bankruptcy and logistic regression, which has a nice probabilistic interpretation because the output is between 0 and 1. However, two major problems with logistic regression occur due to the nature of the bankruptcy problem. First, we assert that bankrupt corporations can be viewed as outliers from the perspective of a group of healthy firms (Booth, 1982). In any given year, the number of corporate bankruptcies is small relative to the total number of publically traded corporations. If bankrupt firms are outliers, this poses a major violation to the underlying distributional assumptions for logistic regression. Outliers in the data set then can lead to inconsistency when maximum likelihood is used to estimate the coefficients for the logistic regression (Bianco and Martinez, 2009). This breakdown of the traditional maximum likelihood (ML) estimator for logistic regression has led to a number of proposals for estimators that are resistant (or robust) to the presence of outliers in the data. The robust estimator we consider is the one due to Bianco and Yohai (1996) as implemented by Croux and Haesbroeck (2003), which we will now refer to as BY robust logistic regression.

The second problem that occurs with logistic regression in the bankruptcy problem is the interpretation of the model percent correctly predicted (or classified). As pointed out by Wooldridge (2009), the percent correctly predicted is a useful goodness-of-fit measure, but it can be very misleading. The percent correctly predicted can be very misleading when the relative ratios of outcomes is large. Again, this is true with corporate bankruptcy because the relative occurrence of bankruptcy is usually very low. Thus in the case of corporate bankruptcy, it is possible to get rather high percentages correctly predicted even when the models' prediction of bankruptcy (the least likely outcome) is extremely poor. Wooldridge (2009) then recommends that researchers compute the percentage correctly predicted for each outcome. Due to the low occurrence of bankruptcy then, most bankruptcy models have relatively poor prediction of bankruptcy.

The main purpose of this study was to investigate the accuracy of predicting bankruptcy using BY robust logistic regression versus ML logistic regression. Using financial ratio data from 2006 and 2007, a three-fold cross validation scheme was designed to compare the correct classification and prediction of bankrupt firms with BY robust logistic regression and ML logistic regression. With both the 2006 and 2007 data, BY robust logistic regression improves both the classification of bankrupt firms in the training set and the prediction of bankrupt firms in the testing set. Our analysis indicates that if the BY robust logistic regression significantly changes the estimated regression coefficients from ML logistic regression, then the BY robust logistic regression method can significantly improve the classification and prediction of bankrupt firms. The reminder of the paper is organized as follows. Section 2 reviews the prior literature on bankruptcy prediction and the development of the BY robust logistic regression method. Section 3 provides the research design and methodologies in terms of data, variables, and the cross-validation scheme. The regression results are shown and discussed in Section 4, and the conclusions of the study are given in Section 5.

2. Literature Review

2.1 Bankruptcy Prediction

Bankruptcy prediction has been a popular subject for business researchers. Beaver (1966) was one of the first researchers to study bankruptcy prediction by testing several financial ratios for their ability to classify and predict bankrupt firms. Altman (1968) introduced bankruptcy models based on discriminant analysis in classifying bankruptcies according to five financial variables: working capital/total assets, retained earnings/total assets, earnings before interest and taxes/total assets, market value of equity/total debt, and sales/total assets. Ohlson (1980) used logistic regression to estimate the probabilities of bankruptcy. Odom and Sharda (1990) were the first researchers to use neural networks for bankruptcy classification and found that neural networks were at least as accurate as discriminant analysis. Since then, a significant volume of neural network research for bankruptcy classification followed (Alam *et al.*, 2000; Jo *et al.*, 1997; Lee *et al.*, 2005; O'Leary, 1998; Tam and Kiang, 1992; Udo, 1993; Wilson and Sharda, 1994; Zhang *et al.*, 1999). Shumway (2001) and Chava and Jarrow (2004) investigated the accuracy of predicting bankruptcy using hazard models.

Bankrupt firms are outliers from the perspective of a group of healthy firms. The fact that only 2% of all firms go bankrupt in normal economic periods suggests (Lenard, Alam and Madey, 1995; O'Leary, 1998) that bankrupt firms can indeed be treated as outliers. While logistic regression provides a nice probabilistic interpretation for bankruptcy, we know that maximum likelihood (ML)

estimates from logistic regression are not resistant to outliers (Bianco and Yohai, 1996). Thus in the next section, we review the development of robust logistic regression methods.

2.2 Bianco-Yohai Robust Logistic Regression

Pregibon (1981) presented a logistic regression analysis of skin vaso-constriction data that contained outliers. His analysis showed that the maximum likelihood (ML) estimates from logistic regression are not robust (i.e. resistant) to outliers. Kunsch *et al.* (1989) then proposed an Optimal Bias-Robust Estimator (OBRE) for generalized linear models. Kunsch *et al.* (1989) presented results for the skin vaso-constriction data, and noted that functions that excessively downweighted outlying observations lead to computational difficulties and that the estimated standard errors increased significantly. Bianco and Yohai (1996) proposed an alternative estimator that was highly robust in the logistic regression model. The Bianco and Yohai (1996) (now referred to as BY) estimator included a bounded function and a bias correction term. Croux and Haesbroeck (2003) proposed a computational method to successfully implement the BY estimator. The Croux and Haesbroeck (2003) procedure uses a bounded function to guarantee the existence of the BY estimator when the ML estimator exists and provides an algorithm to compute the BY estimate. This algorithm is available as a function in R. We utilize this BYLOGREG function to compute the BY logistic regressions in this investigation.

2.2.1 The Bianco-Yohai (BY) Estimator

We consider (following Bianco and Martinez, 2009) a binomial regression model where the response variable Y has a Bernoulli distribution

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}),$$

where F is a strictly increasing cumulative distribution function, $\mathbf{X} \in \mathbb{R}^p$ is a vector of predictor variables and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of unknown regression coefficients. For

$$F(t) = \frac{\exp(t)}{1 + \exp(t)},$$

the logistic regression model obtains. The maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ can be severely affected by outliers. It is known (Bianco and Martinez, 2009) that the MLE breaks down to zero for data sets containing severe outliers. This breakdown behavior has led to a number of proposals for robust estimators of $\boldsymbol{\beta}$. The robust estimator we consider is the one due to Bianco and Yohai (1996) as implemented by Croux and Haesbroeck (2003). Full details of the Bianco

and Yohai (BY) estimator including robustness properties can be found in the referenced papers. Once we have the BY $\widehat{\boldsymbol{\beta}}$, we wish to test hypotheses about the components of $\boldsymbol{\beta}$. Bianco and Martinez (2009) show that using BY estimates in a Wald-type test statistic yields an asymptotic central Chi-square distribution as the test statistic's sampling distribution, just as does the classical Wald statistic in the ML case. Thus for the inference tests, the quadratic form of the Wald-type test statistic reduces to

$$z_i^2 = [\widehat{\boldsymbol{\beta}}_i/\text{Standard error of } \widehat{\boldsymbol{\beta}}_i]^2.$$

Full details of the above are given in Sections 2, 3, 4, and 5 of Bianco and Martinez (2009). The reason BY estimates were used in this study was first, all the estimates are robust to outliers (Bianco and Yohai, 1996). Finally with the recent work of Bianco and Martinez (2009), we can make inference tests since we know the asymptotic distribution of the corresponding Wald-type test statistics.

3. Data and Methodology

The data set, variables used, and the cross-validation scheme are described below.

3.1 Data and Variables

The data sample for this bankruptcy prediction study consists of U.S. corporations that filed for bankruptcy in 2008-2009 as listed in a bankruptcy research database (http://lopucki.law.ucla.edu/corporations.asp). Financial institutions such as commercial banks and investment banks are excluded from the data set because these financial institutions can be affected by actions of government regulators. Furthermore, financial ratio information was required for the bankrupt firms. The financial data was extracted from COMPUSTAT. Originally, the intent was to utilize only firms that filed for bankruptcy in 2008; however, the data availability requirements and exclusion of financial institutions resulted in an insufficient number of bankrupt firms for the cross-validation study. In order to have a sufficient number of bankrupt firms, the sample included firms that filed for bankruptcy during 2009, but prior to June 30, 2009, which resulted in a sample size of 24 bankrupt firms. COMPUSTAT financial data was extracted for the bankrupt firms for 2006 and 2007, which corresponded to the two-year and one-year periods, respectively, prior to bankruptcy filing.

Unlike prior studies such as Lee *et al.* (2005), no matching scheme was used for the non-bankrupt firm sample. The reason that matching was not done in this study is that matching is impossible in "real world" applications where the intent is to predict bankruptcy in a sample of firms. Thus, the non-bankrupt firms for

the data set are then randomly selected from COMPUSTAT firms. Our results show that matching is not required in order to show that BY robust logistic regression improves the classification. Because of the three-fold cross validation scheme chosen for this research, a sample size of 48 non-bankrupt U.S. firms was required. The 48 non-bankrupt firms satisfy the three-fold cross validation scheme and provide sufficient degrees of freedom for the regressions. Financial firms were also excluded from the non-bankrupt sample, and financial ratio data was required for all non-bankrupt firms in the sample. Again, COMPUSTAT financial data was extracted for the non-bankrupt firms for 2006 and 2007.

Each firm was described by Altman's (1968) five financial ratios since the prediction capabilities of these ratios are well documented in the prior literature (Altman, 1968; Bortiz and Kennedy, 1995; Odom and Sharda, 1990; Zhang *et al.*, 1999; Lee *et al.*, 2005):

1. WCTA = working capital / total assets as a measure of the net liquid assets of the firm to total capitalization.

2. RETA = retained earnings / total assets as a measure of cumulative profitability.

3. EBITTA = earnings before interest and taxes / total assets as a measure of the true productivity of the firm's assets.

4. MEDEBT = market value of equity / book value of total debt as a measure of how much the firm's assets can decline in value before the liabilities exceed the assets and the firm becomes insolvent.

5. SALETA = sales / total assets as a measure of the sales generating ability of the firm's assets.

It should be emphasized that we use Altman's (1968) financial ratio model for bankruptcy only as a base model for comparison of the statistical techniques. Again, the main purpose of the study is to compare the accuracy of BY robust logistic regression versus ML logistic regression.

The descriptive statistics of the data set show that on average the WCTA, RETA, and EBITTA ratios were larger for non-bankrupt firms, which indicated that non-bankrupt firms were in a stronger financial condition (The summary statistics are available from the authors upon request). As the firms approached filing for bankruptcy, the WCTA, RETA, and EBITTA ratios all decreased, which indicated the bankrupt firms were in a worsening financial condition as they approached filing for bankruptcy. While the MEDEBT ratio was on average similar between bankrupt and non-bankrupt firms in 2006, the MEDEBT ratio was much lower for the bankrupt firms in 2007. This indicated that in the year

prior to filing for bankruptcy, the firms had lower market values for equity, high total debt, or both.

While the average financial ratios indicated that the data sample properties are consistent with the prior literature, the distributions of some of the financial ratios point to data issues that complicate the statistical analysis. For example, we have stated on average bankrupt firms have lower WCTA, RETA, and EBITTA ratios. However, the firm with the lowest WCTA, RETA, and EBITTA ratios is from the non-bankrupt firm sample. The distribution of the financial ratios is another data issue, especially when one considers the extreme maximum MEDEBT ratio compared to the mean MEDEBT ratio. For example in several cases, the maximum MEDEBT is orders of magnitude larger than the mean, which indicates the presence of outliers in the data sample.

3.2 Cross-Validation Scheme

A cross-validation scheme was developed to investigate the classification performance of logistic regression prediction equations. The cross-validation technique enables us to use the whole data set so that any bias effect would be minimized (Tam and Kiang, 1992; Zhang *et al.*, 1999). In this study a three-fold cross validation technique is used, and Table 1 shows the details of this scheme. The total data set consists of 24 firms that filed for bankruptcy in 2008-2009 and 48 firms that did not file for bankruptcy for a total data set of 72 firms. As shown in Table 1, the total data set is divided into 3 equal and mutually exclusive subsets. Each subset contained 8 randomly selected bankrupt firms and 16 randomly selected non-bankrupt firms for a total of 24 firms. Training was conducted on any two of the three subsets while the remaining subset was used for testing purposes. Thus as can be seen in Table 1, Run 1 uses subset 1 and subset 2 as the training set for the regressions. The prediction equation developed from the training set is then used to predict the probability of bankruptcy for firms in subset 3-the testing data set. The process continues for Run 2 and Run 3 as shown in Table 1 so that each subset is eventually used as the testing data set. The cross-validation scheme was executed separately for the 2006 and 2007 financial data. Finally, the cross-validation scheme shown in Table 1 was then performed using ML logistic regression and BY robust logistic regression.

4. Logistic Regression Results

4.1 Correct Classification and Prediction

In this study, both ML logistic regression and BY robust logistic regression were used to model the probability that a firm filed for bankruptcy with Altman's (1968) five financial ratios as the explanatory variables in the model. Tables 2

Table 1: Three-fold cross-validation technique scheme

	SUBSET			
		1	2	3
RUN	1	Training 16 non-bankrupt 8 bankrupt firms	Training 16 non-bankrupt 8 bankrupt firms	Testing 16 non-bankrupt 8 bankrupt firms
	2	Testing 16 non-bankrupt 8 bankrupt firms	Training 16 non-bankrupt 8 bankrupt firms	Training 16 non-bankrupt 8 bankrupt firms
	3	Training 16 non-bankrupt 8 bankrupt firms	Testing 16 non-bankrupt 8 bankrupt firms	Training 16 non-bankrupt 8 bankrupt firms

Each subset consists of 16 randomly selected non-bankrupt firms and 8 randomly selected bankrupt firms for a total of 24 firms.

Note that in each run, the training set consists of two subsets or 48 total firms and the testing set consists of only one subset or 24 firms.

and 3 summarized the classifications and predictions as applied to the three-fold cross-validation scheme with the 2006 data (Table 2) and 2007 data (Table 3) for both ML logistic regression and BY robust logistic regression. In both tables, each run of the cross-validation scheme shows the number and percentage of the correct classification for the training sets as well as the correct prediction for the testing set. For the purposes of this study, the classification or the prediction of bankruptcy is considered to be correct when the probability of bankruptcy for a bankrupt firm from the regression equation is greater than 0.5.

As discussed previously, correct classification is often difficult with logistic regression because the regression is highly influenced by the relative number of data points in each binary group. Because of this issue, we subdivided the classification and the prediction counts for the non-bankrupt firms and bankrupt firms. Since there are twice as many non-bankrupt firms in the sample, we expect the regression model to be pulled toward the non-bankrupt firms. In fact, we see in the training sets that the correct classification for non-bankrupt firms is over 93% for ML regression for both 2006 data (Table 2) and 2007 data (Table 3). Since there are less bankrupt firms, we expect the regression equation to be pulled away from bankrupt firms, and thus the correct classification percentage for bankrupt firms should fall. Indeed, we see the correct classification of bankrupt firms with ML regression in 2006 (Table 2) averages 21% in the 3 runs. In Table 3 with the 2007 data, the firms move closer to bankruptcy and we expect the classification of bankrupt firms with ML regression to improve. As expected, the correct classification of bankrupt firms with the 2007 data improves from 21%

Table 2: Comparison of correct classifications and predictions for ML logistic regression versus BY logistic regression for 2006 financial data

ML Logistic Regression						
	Training Set			Testing Set		
	Non-bankrupt	Bankrupt	Overall	Non-bankrupt	Bankrupt	Overall
Run 1						
Correct #	30	6	36	16	2	18
Total #	32	16	48	16	8	24
Percent Correct	93.75%	37.50%	75.00%	100.00%	25.00%	75%
Run 2						
Correct #	31	0	31	15	1	16
Total #	32	16	48	16	8	24
Percent Correct	96.88%	0.00%	64.58%	93.75%	12.50%	66.67%
Run 3						
Correct #	30	4	34	14	1	15
Total #	32	16	48	16	8	24
Percent Correct	93.75%	25.00%	70.83%	87.50%	12.50%	62.50%
Summary						
Correct #		10	101		4	49
Total #		48	144		24	72
Percent Correct		20.83%	70.14%		16.67%	68.06%
BY Logistic Regression						
	Training Set			Testing Set		
	Non-bankrupt	Bankrupt	Overall	Non-bankrupt	Bankrupt	Overall
Run 1						
Correct #	30	14	44	14	5	19
Total #	32	16	48	16	8	24
Percent Correct	93.75%	87.50%	91.67%	87.50%	62.50%	79.17%
Run 2						
Correct #	31	0	31	15	1	16
Total #	32	16	48	16	8	24
Percent Correct	96.88%	0.00%	64.58%	93.75%	12.50%	66.67%
Run 3						
Correct #	30	4	34	14	1	15
Total #	32	16	48	16	8	24
Percent Correct	93.75%	25.00%	70.83%	87.50%	12.50%	62.50%
Summary						
Correct #		18	109		7	50
Total #		48	144		24	72
Percent Correct		37.50%	75.69%		29.17%	69.44%

Table 3: Comparison of correct classifications and predictions for ML logistic regression versus BY logistic regression for 2007 financial data

ML Logistic Regression						
	Training Set			Testing Set		
	Non-bankrupt	Bankrupt	Overall	Non-bankrupt	Bankrupt	Overall
Run 1						
Correct #	30	9	39	12	2	14
Total #	32	16	48	16	8	24
Percent Correct	93.75%	56.25%	81.25%	75.00%	25.00%	58.33%
Run 2						
Correct #	32	0	32	16	0	16
Total #	32	16	48	16	8	24
Percent Correct	100.00%	0.00%	66.67%	100.00%	0.00%	66.67%
Run 3						
Correct #	30	6	36	12	1	13
Total #	32	16	48	16	8	24
Percent Correct	93.75%	37.50%	75.00%	75.00%	12.50%	54.17%
Summary						
Correct #		15	107		3	43
Total #		48	144		24	72
Percent Correct		31.25%	74.31%		12.50%	59.72%
BY Logistic Regression						
	Training Set			Testing Set		
	Non-bankrupt	Bankrupt	Overall	Non-bankrupt	Bankrupt	Overall
Run 1						
Correct #	28	10	38	12	4	16
Total #	32	16	48	16	8	24
Percent Correct	87.50%	62.50%	79.17%	75.00%	50.00%	66.67%
Run 2						
Correct #	28	11	39	14	3	17
Total #	32	16	48	16	8	24
Percent Correct	87.50%	68.75%	81.25%	87.50%	37.50%	70.83%
Run 3						
Correct #	30	6	36	12	1	13
Total #	32	16	48	16	8	24
Percent Correct	93.75%	37.50%	75.00%	75.00%	12.50%	54.17%
Summary						
Correct #		27	113		8	46
Total #		48	144		24	72
Percent Correct		56.25%	78.47%		33.33%	63.89%

correct to 31% correct. At this point, it is important to emphasize the difficulty of predicting bankrupt firms. Although there are only two possible outcomes, it is always easier to predict non-bankrupt firms since non-bankrupt firms are more numerous in the data sets and in the "real" world. In all the regressions, the correct classification of non-bankrupt firms is about 90%. The difficulty is classifying and predicting the bankrupt firms since they are essentially outliers.

With BY robust logistic regression, again our focus is on the classification of bankrupt firms. From Table 2, we see that the correct classification of bankrupt firms with the 2006 data is 37.5%. Based on the 2006 financial data, we find that BY robust logistic regression improved the classification of bankrupt firms over ML logistic regression from 21% to 37.5% correct. In Table 3 with the 2007 financial data, we see again that BY robust regression improved the classification of bankrupt firms over ML regression from 31% to 56% correct.

With BY robust logistic regression, the improvement in classification of the bankrupt firms also translates to improvement in the prediction of bankrupt firms in the testing set. In the 2006 data (Table 2), the BY robust logistic regression improved the prediction of bankrupt firms over ML logistic regression from 16.7% to 29.2% correct. BY robust logistic regression improved the prediction of bankrupt firms in the 2007 data sample (Table 3) over ML logistic regression from 12.5% to 33.3%. With the three-fold cross-validation study, BY robust logistic regression improved the classification of bankrupt firms in the training sets. Furthermore, BY robust logistic regression also improved the prediction of bankrupt firms in the testing sets. In fact, on an overall classification and prediction basis, the BY robust logistic regression was superior to the ML logistic regression. The contribution of this research is to show that BY robust logistic regression improves the classification and prediction of bankrupt firms over ML logistic regression. While there are more sophisticated logistic regression bankruptcy models than Altman's (1968) financial ratio model that could be used to improve the classification, our main point is that BY robust logistic regression is a better statistical technique for the classification and prediction of bankrupt firms for a given model.

4.2 Analysis of Regression Coefficients

In order to develop some insights into the superior classification and prediction results of the BY robust logistic regression, it was useful to examine the regression coefficients estimated from the training sets. The estimated regression coefficients for the ML logistic regression and BY robust logistic regression were summarized in Table 4, with the results from the 2006 data shown in Panel A and the results from the 2007 data shown in Panel B. In reviewing the 2006 data in Panel A of Table 4, we note that the regression coefficients estimated from the BY robust logistic regression are significantly different from the coefficients estimated from

Table 4: Summary of logistic regression results

PANEL A 2006 Financial Data

	Parameter Estimates (Standard errors below in parentheses)					
	ML Logistic Regression			BY Logistic Regression		
	RUN 1	RUN 2	RUN 3	RUN 1	RUN 2	RUN 3
Variable						
Intercept	-0.2387	-0.4033	-0.6379	1.4597	-0.4014	-0.6079*
	(.7219)	(.5832)	(.6253)	(.9489)	(.2998)	(.3426)
WCTA	0.0603	-1.7855	-1.8254	22.0113**	-1.7773*	-1.7395*
	(1.5791)	(1.7376)	(1.9722)	(10.6710)	(.9333)	(.9656)
RETA	0.0762	0.2544	0.8465	-0.5861	0.2532**	0.8067***
	(.1416)	(.2240)	(.6722)	(.3172)	(.1103)	(.2971)
EBITTA	-0.0165	-0.2476	-1.6378	-1.4556	-0.2465	-1.5607*
	(1.3017)	(1.5305)	(2.1121)	(1.0454)	(.7112)	(.9113)
MEDEBT	-0.0697	-0.0006	-0.0006	-1.8140**	-0.0006	-0.0006
	(.0584)	(.0013)	(.0015)	(.74122)	(.0005)	(.0005)
SALETA	0.2587	0.1734	0.5077	0.7750	0.1727	0.4838*
	(.4913)	(.4189)	(.4818)	(.5796)	(.2300)	(.2635)

PANEL B 2007 Financial Data

	Parameter Estimates (Standard errors below in parentheses)					
	ML Logistic Regression			BY Logistic Regression		
	RUN 1	RUN 2	RUN 3	RUN 1	RUN 2	RUN 3
Variable						
Intercept	-0.1635	-0.4183	-0.4915	0.9542	0.4508	-0.4733
	(.7233)	(.5592)	(.6087)	(.6520)	(.3509)	(.3150)
WCTA	-0.9024	-0.0858	-2.0125	-2.9568**	3.1891	-1.9381**
	(.7520)	(.6510)	(1.7516)	(1.4804)	(2.0803)	(.8633)
RETA	0.0827	0.0580	0.8992	0.1301**	-0.4185*	0.8659***
	(.1288)	(.1697)	(.6650)	(.0636)	(.2383)	(.3303)
EBITTA	0.9285	-0.0315	-0.9785	1.8078**	1.1725	-0.9423
	(1.2619)	(1.1177)	(1.9136)	(.7709)	(1.2112)	(.9276)
MEDEBT	-0.1543	-0.0208	-0.0162	-0.6442*	-1.0838*	-0.0156
	(.0903)*	(.0212)	(.0240)	(.3357)	(.5699)	(.0128)
SALETA	0.3658	0.0956	0.4447	0.1441	0.3329	0.4282**
	(.5262)	(.3707)	(.4474)	(.3217)	(.3247)	(.2126)

* indicates asymptotic significance at 10% level
** indicates asymptotic significance at 5% level
*** indicates asymptotic significance at 1% level

the ML logistic regression for Run 1. A review of the classification and prediction results for bankrupt firms from Table 2 indicates that BY robust logistic regression was also far superior to ML logistic regression for Run 1. However, if we examine Runs 2 and 3 in Table 2, we note that BY robust logistic regression does not improve the classification and prediction results over ML logistic regression. In reviewing the estimated regression coefficients in Panel A of Table 4 for Runs 2

and 3, we note that the estimated regression coefficients are essentially the same for ML logistic regression and BY robust logistic regression.

A similar pattern can be found in Panel B of Table 4 for the estimated coefficients from the 2007 data. In Runs 1 and 2 in Panel B of Table 4, the regression coefficients estimated from BY robust logistic regression are significantly different from the coefficients estimated from ML logistic regression for these cases. A review of the classification and prediction results for bankrupt firms from Table 3 indicates that BY robust logistic regression was again superior to ML logistic regression for Runs 1 and 2. Meanwhile, in Run 3 in Panel B of Table 4, the estimated regression coefficients are essentially the same for ML logistic regression and BY robust logistic regression. Then a review of Run 3 in Table 3 indicates that BY robust logistic regression does not improve the classification and prediction results over ML logistic regression. This analysis indicates that if the BY robust logistic regression significantly changes the estimated regression coefficients from ML logistic regression, then the BY robust logistic regression method can significantly improve the classification and prediction of bankrupt firms. At worst, the BY robust logistic regression makes no changes in the estimated regression coefficients and has the same classification and prediction results as ML logistic regression. This is strong evidence that BY robust logistic regression should be used as a robustness check on ML logistic regression, as well as for prediction when outliers exist in the data set.

Another reason that BY robust logistic regression should be used as a robustness check on ML logistic regression is that BY robust logistic regression can provide different standard errors. Hauser and Booth (2011) showed that BY robust logistic regression can lead to different interpretations on the significance of explanatory variables, even if the estimated coefficients are similar. Indeed in Table 4, we can see that BY robust logistic regression yields different estimated standard errors, which consequently can lead to different interpretations. Consider that in all the runs of the three-fold cross-validation scheme, ML logistic regressions found only 1 of the Altman (1968) financial ratios significant at the 10% level in 1 run and no variables significant at the 5% level in any case. As can be seen in Table 4, BY robust logistic regression found more than one of the Altman (1968) financial ratios was significant at the 10% level in every run of the three-fold cross validation scheme. We have stated that in Run 3 in Panel B of Table 4 the estimated regression coefficients are essentially the same for ML logistic regression and BY robust logistic regression. Also in Run 3, BY robust logistic regression made no improvement in classification or prediction of bankrupt firms. However, BY robust regression found the financial variables WCTA, RETA, and SALETA to be significant at the 5% level in Run 3 while ML logistic regression showed no significant explanatory variable even at the 10% level in Run 3.

The use of BY robust logistic regression, then provides the researcher with another "tool" to analyze the ML regression results. That is, if the results are the same with BY robust logistic regression and ML logistic regression, improvement in the classification accuracy can only be achieved by improvements to the model such as including additional variables.

4.3 Bankruptcy Prediction of Lehman Brothers

While the three-fold cross-validation scheme showed that BY robust logistic regression provided superior classification and prediction of bankrupt firms, it was interesting to examine the models' prediction of the Lehman Brothers bankruptcy in 2008. The Lehman Brothers bankruptcy provided a case study that was clearly outside of the data set since it was a failed investment bank, but the bankruptcy was not forced by regulators. The five Altman (1968) financial ratios were computed from COMPUSTAT and SEC 10-K filings (http://www.sec.gov/Archives/edgar/data/806085/000110465908005476/a08-3530_110k.htm) for 2006 data and 2007 data. Then using the regression equations from the Runs 1-3 training sets, the probability of bankruptcy was computed using the Lehman Brothers financial ratios. These calculations were summarized in Table 5. From Table 5, we can see that regression equations from the ML logistic regression never predicted bankruptcy with either the 2006 or 2007 data. However, the BY robust logistic regression predicted bankruptcy for Lehman Brothers, though not with the prediction equation from every run. With the 2006 data, the regression equation from Run 1 predicted bankruptcy of Lehman Brothers while Run 1 and Run 2 regression equations with the 2007 data predicted bankruptcy for Lehman Brothers. It is not surprising that these are the same runs where BY robust logistic regression improved the classification of bankrupt firms and had significantly different estimated coefficients from ML logistic regression. These calculations show again that in cases where the BY robust logistic regression significantly changes the estimated coefficients from ML logistic regression, BY robust logistic regression improves the prediction of bankrupt firms. No other "out of sample" analysis or data extrapolation was done with the BY robust logistic regression.

4.4 Analysis of Deviance Residuals

Given that the classification or prediction of bankrupt firms can be considered a problem in outlier detection (from the perspective of a group of firms) and that the BY robust logistic regression is resistant to the presence of outliers, it seems that BY robust logistic regression should improve the prediction of bankrupt firms over ML logistic regression. The results of this study confirmed and quantified this hypothesis. In this section, we address the issue of outliers in the sample.

Table 5: Bankruptcy prediction for Lehman Brothers holdings, Inc.

In this Table, the prediction equations from the Run 1, Run 2, and Run 3 training set regressions were used to predict the probability of bankruptcy for Lehman Brothers, an out of sample and out of model case.

PANEL A 2006 Financial Data

	Bankruptcy Prediction					
	ML Logistic Regression			BY Logistic Regression		
	RUN 1	RUN 2	RUN 3	RUN 1	RUN 2	RUN 3
Probability	0.448	0.329	0.268	0.994	0.33	0.278
Correct Prediction #	0	0	0	1	0	0

OVERALL	ML Logistic Regression	BY Logistic Regression
Correct Prediction %	0.0%	33.3%

PANEL B 2007 Financial Data

	Bankruptcy Prediction					
	ML Logistic Regression			BY Logistic Regression		
	RUN 1	RUN 2	RUN 3	RUN 1	RUN 2	RUN 3
Probability	0.452	0.396	0.32	0.661	0.708	0.326
Correct Prediction #	0	0	0	1	1	0

OVERALL	ML Logistic Regression	BY Logistic Regression
Correct Prediction %	0.0%	66.7%

Note that predicted probabilities greater than 0.5 were considered correct, and less than 0.5 were considered incorrect.

On a univariate basis, we discussed the distributions of some of the financial ratios used as explanatory variables in the logistic regressions. For example, recall that in several cases the maximum MEDEBT ratio is several orders of magnitude larger than the mean indicating the presence of univariate outliers in the data set. Since we are interested in predicting bankruptcy however, we are more interested in multivariate outliers, which then are more complicated to detect because of the complex nature of bankruptcy. An indication of the presence of multivariate outliers can be seen from an analysis of the deviance residuals

from the ML logistic regression and the properties of the deviance residuals. The maximum deviance residual is largest with both the 2006 and 2007 data in Run 1 of the three-fold cross validation scheme. Since the maximum deviance residual is relatively large, we would interpret this to indicate the presence of at least 1 multivariate outlier. In the presence of outliers, we expect the BY robust logistic regression to outperform ML logistic regression. This is indeed the case for Run 1 with both 2006 and 2007 data. A large deviance residual in these cases seems to indicate outliers, and the outlier resistant BY robust logistic regression yields significantly different estimated coefficients and better prediction of bankrupt firms. The maximum deviance residual is smallest with both the 2006 and 2007 data in Run 3. In these cases, we would interpret the relatively small deviance residuals to indicate no significant outliers. Without significant outliers, BY robust logistic regression would yield essentially the same results as ML logistic regression. Indeed, we see essentially the same results for BY robust logistic regression and ML logistic regression in Run 3 for both the 2006 and 2007 data.

We should emphasize at this point that this analysis of the maximum deviance residual (and its distribution properties) is only an indication of the influence of multivariate outliers. Future research is required to better define test statistics, which would better define the influence of multivariate outliers. In lieu of such a test statistic that defines the influence of outliers, our results indicate that BY robust logistic regression should be done as a robustness check on the ML logistic regression. If there are outliers in the data sample, the BY robust logistic regression will result in significantly different estimated coefficients and better bankruptcy prediction. If there are no significant outliers in the data sample, the BY robust logistic regression will produce essentially the same results as ML logistic regression.

5. Conclusions

The main purpose of this study was to investigate the accuracy of predicting bankruptcy using BY robust logistic regression versus ML logistic regression. The data set for the study was a sample of 24 non-financial U.S. firms that filed for bankruptcy in 2008-2009 and a sample of 48 non-financial U.S. firms that did not file for bankruptcy in 2008-2009. Using financial ratio data from 2006 and 2007, a three-fold cross validation scheme was designed to compare the correct classification and prediction of bankrupt firms with BY robust logistic regression and ML logistic regression.

With both the 2006 and 2007 data, BY robust logistic regression improved both the classification of bankrupt firms in the training set and the prediction of bankrupt firms in the testing set. In the 2006 data, the BY robust logistic regression improved the prediction of bankrupt firms over ML logistic regression from

16.7% to 29.2% correct. BY robust logistic regression improved the prediction of bankrupt firms in the 2007 data sample over ML logistic regression from 12.5% to 33.3% correct. On an overall classification and prediction basis, the BY robust logistic regression was superior to the ML logistic regression. While there are more sophisticated logistic regression bankruptcy models than Altman's (1968) financial ratio model that could be used to improve the classification, our main point is that BY robust logistic regression is a better statistical technique for the classification and prediction of bankrupt firms for a given model.

In an out of sample case study with the failed investment bank Lehman Brothers, the ML logistic regression never predicts bankruptcy with either the 2006 or 2007 data. However, the BY robust logistic regression was able to correctly predict bankruptcy for Lehman Brothers.

A review of the estimated coefficients from BY robust logistic regression indicated that improved classification and prediction of bankrupt firms occurred when the estimated coefficients from BY robust logistic regression were significantly different from the coefficients estimated from ML logistic regression. Since the BY robust logistic regression is robust to the presence of outliers, we showed evidence that BY robust logistic regression improves on the ML logistic regression when outliers are present in the sample.

Our analysis indicates that if the BY robust logistic regression significantly changes the estimated regression coefficients from ML logistic regression, then the BY robust logistic regression method can significantly improve the classification and prediction of bankrupt firms. At worst, the BY robust logistic regression makes no changes in the estimated regression coefficients and has the same classification and prediction results as ML logistic regression. This is strong evidence that BY robust logistic regression should be used as a robustness check on ML logistic regression. If a difference exists, BY robust logistic regression should be used as the primary classifier of bankrupt firms.

Acknowledgements

We thank the editor and reviewers for their very helpful advice. We also thank Dr. Croux and Dr. Haesbroeck for the use of the R function, BYLOGREG.

References

Alam, P., Booth D., Lee, K. and Thordarson, T. (2000). The use of fuzzy clustering algorithm and self organizing neural networks for identifying potentially failing banks: an experimental study. *Expert Systems with Applications* **18**, 185-199.

Altman, E. L. (1968). Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. *Journal of Finance* **23**, 589-609.

Beaver, W. (1966). Financial ratios as predictors of failure, empirical research in accounting, selected studies 1966. *Journal of Accounting Research* **4**, 71-111.

Bianco, A. M. and Martinez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics and Data Analysis* **53**, 4095-4105.

Bianco, A. M. and Yohai, V. J. (1996). *Robust Estimation in the Logistic Model, in Robust Statistics, Data Analysis, and Computer Intensive Methods*, 17-34; *Lecture Notes in Statistics* **109**, Ed. H. Rieder, Springer Verlag, New York.

Booth, D. (1982). The analysis of outlying data points by robust regression: A multivariate problem bank identification model. *Decision Sciences* **13**, 72-81.

Boritz, J. E. and Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications* **9**, 503-512.

Chava, S. and Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance* **8**, 537-569.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for Logistic Regression. *Computing Statistical and Data Analysis* **44**, 273-295.

Hauser, R. and Booth D. (2011). CEO bonuses as studied by robust logistic regression. *Journal of Data Science* **9**, 293-310.

Kordzakhia, N., Mishra, G. D. and Reiersolmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference* **98**, 211-223.

Kunsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989). Conditionally unbiased bounded Influence estimation in general regression models with applications to generalized linear models. *Journal of the American Statistical Association* **84**, 460-466.

Jo, H.Y., Han, I. G. and Lee, H. Y. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications* **13**, 97-108.

Lee, K., Booth, D. and Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy in Korean firms. *Expert Systems with Applications* **29**, 1-16.

Lenard, M. J., Alam, P. and Madey, G. R. (1995). The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. *Decision Sciences* **26**, 209-227.

Odom, M. and Sharda, R. (1990). A neural network model for bankruptcy prediction. *Proceedings of the IEEE International Conference on Neural Networks* **2**, 163-168.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* **18**, 109-131.

O'Leary, D. E. (1998). Using neural networks to predict corporate failure. *International Journal of Intelligent Systems in Accounting, Finance, and Management* **7**, 187-197.

Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics* **9**, 705-724.

Sharda, R. and Wilson, R. L. (1996). Neural network experiments in business-failure forecasting: predictive performance measurement issues. *International Journal of Computational Intelligence and Organization* **1**, 107-117.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business* **74**, 101-124.

Tam, K. Y. and Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management Science* **38**, 926-947.

Wilson, R. L. and Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems* **11**, 545-557.

Wooldridge, J. (2009). *Introductory Econometrics 4th ed.*, South-Western Cengage Learning.

Zhang, G., Hu, M. Y., Patuwo, B. E. and Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. *European Journal of Operational Research* **116**, 16-32.

Received June 17, 2010; accepted December 20, 2010.

Richard P. Hauser
Department of Finance
Kent State University
P.O. Box 5190 Kent, OH 44242-0001, USA
rhauser1@kent.edu

David Booth
Department of Management and Information Systems
Kent State University
P.O. Box 5190 Kent, OH 44242-0001, USA
dbooth@kent.edu