Chapter 62
The Prediction of Default with Outliers: Robust Logistic Regression

Chung-Hua Shen, Yi-Kai Chen, and Bor-Yi Huang

Abstract This paper suggests a Robust Logit method, which extends the conventional logit model by taking outliers into account, to implement forecast of defaulted firms. We employ five validation tests to assess the in-sample and out-of-sample forecast performances, respectively. With respect to in-sample forecasts, our Robust Logit method is substantially superior to the logit method when employing all validation tools. With respect to the out-of-sample forecasts, the superiority of Robust Logit is less pronounced.

Keywords Logit • Robust Logit • Forecast • Validation test

62.1 Introduction

In recent years a large number of researchers and practitioners have worked on the prediction of business defaults. This prediction is important because not only can it reduce nonperforming loans but it can help to determine capital allocation. As required by the Basel Committee on Banking Supervision, the prediction of business default is the first step to fulfill the requirement of the internal rating-based (IRB) of Basel II. Large banks are therefore eager to develop their default prediction system to monitor credit risk. One of the issues regarding credit risk assessment is the model or method of default prediction used. Altman and Saunders (1998) have traced development of the literatures in risk measurement for 20 years. Currently, the common model used includes discriminant analysis (Altman 1968), logit and probit models (Ohlson 1980; Westgaard and Wijst 2001), multi-group hierarchy model (Doumpos et al. 2002; Doumpos and Zopounidis 2002), neural network (Atiya 2001; Piramuthu 1999; Wu and Wang 2000) and option type models such as KMV to name a few. The works of Lennox (1999) for the comparison of the first three models and Dimitras et al. (1996) for the discussion of the first five models have become the standard reference for the financial research.

While there are many methods to estimate the probability of default, none of them have taken the outliers into account when there is a discrete dependent variable. Outliers that can seriously distort the estimated results have been well documented in the conventional regression model. For example, Levine and Zervos (1998) employ 47 countries data and confirm that the liquidity trading is positively related to economic growth. Zhu et al. (2002), however, reject this positive relation when they employ the econometric methods to minimize the outlier distortion effects caused by Taiwan and South Korea. Although methods and applications that take outliers into account are well known when the dependent variables are continuous (Rousseeuw 1984; Rousseeuw and Yohai 1984), few have conducted empirical studies when the dependent variable is binary. Atkinson and Riani (2001), Rousseeuw and Christmann (2003), and Flores and Garrido (2001) have developed the theoretical foundations as well as the algorithm to obtain consistent estimator in logit models with outliers, but they do not provide applied studies. If outliers indeed exist when the dependent variable is binary, the conventional logit model might be biased.

The aim of this paper is to predict default probability with the consideration of outliers. This is a direct extension of the logit estimation method and is referred to as the Robust Logit model hereafter. We apply the forward search method of Atkinson and Cheng (2000) and Atkinson and Riani (2001) to Taiwan data. To the best of our knowledge, our paper is the first to use the Robust Logit model for actual data. Once estimated coefficients are obtained, we assess the performances of the logit and the Robust Logit methods by using five validation tools; that is, contingency table (cross-classifications table), cumulative accuracy profile (CAP), relative or receive operation characteristics (ROC), Kolmogorov-Smirnov (KS), and Brier score.

C.-H. Shen (✉)
Department of Finance, National Taiwan University, Taipei, Taiwan
e-mail: chshen01@ntu.edu.tw

Y.-K. Chen
Department of Finance, National University of Kaohsiung, Kaohsiung, Taiwan
e-mail: chen@nuk.edu.tw

B.-Y. Huang
Department of Business Management, Shih Chien University, Taipei, Taiwan
e-mail: hby-1688@mail.usc.edu.tw

C.-F. Lee et al. (eds.), *Handbook of Quantitative Finance and Risk Management*,
DOI 10.1007/978-0-387-77117-5_62,

The paper proceeds as follows. In addition to the first section, the next section provides literatures of the logit and the Roubst Logit regression model. Section 62.3 introduces the five validation models. Section 62.4 discusses the empirical model and estimated results. Section 62.5 provides the conclusion.

62.2 Literature Review of Outliers in Conventional and in Logit Regression

Conventional linear regression analysis taking outliers into account has been utilized since 1960. The methods, such as Least Median of Squares (LMS), Least Trimmed Squares (LTS) (Rousseeuw 1983, 1984), which exclude the effect of outliers on linear regression, are now standard options in many econometric softwares. Until 1990, however, the literature was slow in the consideration of outliers when the logit model is involved. Furthermore, most development tends to center on the theoretical derivations of outliers in logit method in the fields of statistics and actuaries.

62.2.1 Outliers in Conventional Regression

Cook and Weisberg (1982) suggest a backward procedure to exclude outliers in the conventional linear model. By using the whole sample as the first step, Cook and Weisberg (1982) detect one outlier and remove it. Then, they go on detecting the second outlier and removing it, followed by the third and so on. Repeating this step, they remove all outliers. While the backward approach appears straightforward, it, however, has suffered from a masking effect. Namely, the statistical properties of the detected outliers, are affected by the outliers remaining in the sample. Barrett and Gray (1997) and Haslett (1999) suggest a multiple outliers' method to overcome this problem but the speed of detection is slow. Atkinson (1985, 1994) proposes a forward search algorithm (forward approach), which uses only a small subset without outliers as the first step. Then, another subset is added and examine in the next observation. Continuing this step, he claims that the forward approach can remove outliers without the masking effect. When the number of outliers is not known, Atkinson and Riani (2006) indicate a method that makes efficient use of the individual simulations to derive the simultaneous properties of the series of tests occurring in the practical data-analytical case. Riani and Atkinson (2007) provide easily calculated bounds for the statistic during the forward search and illustrate the importance of the bounds in inference about outliers.

62.2.2 Outliers in Logit Regression: Robust Logistic Regression

Our Robust Logistic (RL) regression is based on Atkinson and Riani (2001) forward approach, which include five steps.

<*Step 1*> Choice of the initial subset of observations

Randomly choose $k + 1$ observations where $k + 1$ is equal to one-third of total observations as our starting sample size.[1] The corresponding estimated coefficient vector of logit method is denoted as $\hat{\beta}^{(\mathrm{k}+1)}$ and the predicted value of the observed company is $\hat{y}_i = F\left(x_i\hat{\beta}^{(k+1)}\right)$, where $i = 1\ldots, N$.

<*Step 2*> Obtain the median of the errors

Calculate the probability of accurate rate of the prediction of the default companies as $p^{(k+1),i}$:

$$p^{(k+1),i}\begin{cases} = \hat{y}_i,\ \text{if } y_i = 1 \\ = 1 - \hat{y}_i,\ \text{if } y_i = 0. \end{cases}$$

Corresponding to the accurate rate of the prediction, the probability of the nonaccurate rate of the prediction $e^{(k+1),i}$ is calculated as $e^{(k+1),i} = 1 - p^{(k+1),i}$ of all observations. Then take an ascending order of all $e^{(k+1),i}$, i.e.,

$$e^{(k+1),1} < e^{(k+1),2} < \cdots < e^{(k+1),N},$$

and obtain the median of all $e^{(k+1),i}$ which is $e^{(k+1),med}$.

<*Step 3*> Proceed forward search algorithm

Add an additional observation in the subset. Employing $k + 2$ observations to yield coefficients $\hat{\beta}^{(\mathrm{k}+2)}$. Yet, these $k + 2$ observations are the observations corresponding to the smallest errors of $(k + 2)$ observations in *Step 2*; that is, observations corresponding to $e^{(k+1),1}, e^{(k+1),2}, \ldots, e^{(k+1),k+2}$. This is equivalent to removing the outliers.

Then, repeating *Step 2* and we can obtain the median of $e^{(k+2),i}$, which is $e^{(k+2),med}$.

<*Step 4*> Obtain all estimated coefficients and corresponding error median

Add an additional observation again. It means that repeat *Step 3* by adding another observation and use $k + 3$ observations corresponding to the smallest k+3 errors of $e^{(k+2),i}$ in *Step 3* We then similarly obtain $\hat{\beta}^{(\mathrm{k}+3)}$ and median $e^{(k+3),med}$.

Repeat the above steps by adding one additional observation in each estimation and the process is done until all samples are used. We thus obtain estimated coefficients $\hat{\beta}^{(\mathrm{k}+4)}, \hat{\beta}^{(\mathrm{k}+5)}, \ldots, \hat{\beta}^{\mathrm{N}}$ and the corresponding median $e^{(k+4),med}, e^{(k+5),med}, \ldots, e^{N,med}$.

<*Step 5*> Outlier is found

[1] Atkinson uses $k = +1$ as the number of parameters $+1$ as the starting sample size. We do not adopt his suggestion because the small sample size often is full of zeros without one, invalidating the logit model.

Calculating $e^{*,med} = \min\left[e^{(k+1),med}, e^{(k+2),med}, \ldots e^{N,med}\right]$ and its corresponding $\hat{\beta}^{*}$, which is the estimator of RL method.[2]

Although this forward search method is intuitively appealing, it encounters three problems in the actual application. First, the random sampling may pick all zeros or ones as dependent variable, which fail the estimation. Next, the selected initial set of samples affects the estimation results. Thus, repeated sampling of the initial set become necessary. Third, companies identified as outliers may be those extremely good and bad companies. They are statistical outliers but not financial outliers.

62.3 Five Validation Tests

Once we obtain the estimated results from two methods, we compare their forecasting ability based on the following five validation tests for the assessment of discriminatory power. The validation methods introduced here are mostly based on the work by Sobehart and Keenan (2004), Sobehart et al. (2000) and Stein (2002).

62.3.1 Contingency Table (Cross-Classification Table)

Contingency Table, also referred as the Cross-Classification Table, is the most often used validation tool in comparing the power of prediction. Let *TP%* and *TN%* be the ratios of success in predicting default and non-default firms, whereas *FP%* and *FN%* be the ratios of failure in predicting default and non-default firms (see Table 62.1). In conventional terms, the sum of *TP%* and *TN%* is referred to as the hit rate, whereas *FP%* and *FN%* are referred to as Type II and Type I errors, respectively. Furthermore, $TP\% + FN\% = 1$ and $FP\% + TN\% = 1$. The weakness of this table is that only one cutoff is chosen to decide these ratios. Typically, the selection of this cutoff is based on the average rule (the cutoff is then 0.5) or the sample proportion rule (the cutoff is then the number of default/total number firms). More cutoffs may be needed, which motivates the development of the following validation tests.

Table 62.1 Contingency table (cross-classification table)

Actual \ Predicted	Default companies ($y_i = 1$)	Non-default companies ($y_i = 0$)
Default companies ($\hat{y}_i = F\left(x_i\hat{\beta}^*\right) \geq$ cutoff)	$TP\% \times D$	$FP\% \times N$
Non-default companies ($\hat{y}_i = F\left(x_i\hat{\beta}^*\right) <$ cutoff)	$FN\% \times D$	$TN\% \times N$

Notations are taken from Sobehart et al. (2000). *TP* true positive means that companies are default and are accurately predicted; *FN* false negative means companies are not default and not correctly predicted; *FP* false positive means companies are default and not correctly predicted; *TN* true negative means that companies are not default and correctly predicted; *D* is number of default and *N* is number of non-default companies

62.3.2 Cumulative Accuracy Profile (CAP)

CAP curve is a visual tool which graph can easily be drawn if two representative samples of scores for defaulted and non-defaulted borrowers are available. The shape of the concavity of the CAP is equivalent to the property that the conditional probabilities of default, given the underlying scores, form a decreasing function of the scores (default probability). Alternatively, non-concavity indicates suboptimal use of information in the specification of the score function. Researchers typically calculate the accuracy ratio, which is the area under the rating model divided by the area under the perfect model, to examine the performance of model. This is equivalent to A/B graphed in Fig. 62.1.

Figure 62.1 plots CAPs of a perfect rating and random model. A perfect rating model will assign the lower estimated scores to the defaulters. In this case the CAP is increasing linearly and then staying at one. For a random model without any discriminative power, the fraction x of all debtors with the lower scores contain x percent of all defaults. Applied rating systems will be somewhere in between these two extremes. Statistically, the comparison ratio is defined as the

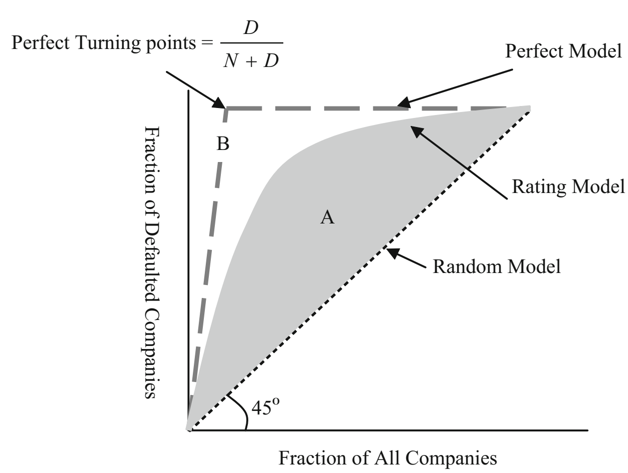

Fig. 62.1 CAP curve

[2] We could further repeat *Step 1* to start different set of observations.

ratio of A/B, where A is the area between the CAP of the rating model being validated and the CAP of the random model, and B is the area between the CAP of the perfect rating model and the CAP of the random model.

The calculation of area under rating model is as follows. First, a descending order of the estimated default rates is ranked. Then, it takes the top $s\%$ number of firms that have the higher estimated default rates, making these numbers equal to $G = s\% \times (N + D)$ where N and D are the number of non-defaulting and default companies in the data set. Within G firms, it then calculates the number of firms that are actually default and are divided by G to yield $y\%$. Repeating the above process, we obtain a sequence of $s\%$ and $y\%$. Plotting $y\%$ (y-axis) against $s\%$ (x-axis) yields CAP. The shape of the rating model depends on the proportion of solvent and insolvent borrowers in the sample.

62.3.3 Receiver Operating Characteristic (ROC)

ROC curve uses the same information as CAP to answer the question: What percentage of non-defaulters would a model have to exclude? (Stein 2002). It generalizes Contingency Table analysis by providing information on the performance of a model at any cutoff that might be chosen. It plots the *FP%* rate against the *TP%* rate for all credits in a portfolio. In particular, ROCs are constructed by scoring all credits and ordering the non-defaults from the worst to the best on the x axis and then plotting the percentage of defaults excluded at each level on the y axis. The area under the rating model is

$$\mathrm{A} + \mathrm{B} = \sum_{k=0}^{k=n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k) \div 2,$$

where A and B are the area under rating model and 45°, respectively and n is the number of intervals. The ROC curve is demonstrated in Fig 62.2.

For example, assuming that the number of default firm $D = 50$, and non-default firm $N = 450$. Then, similar to the CAP method, a descending order of the estimated default probability is ranked. Next, giving a fixed type II error, *FP%*, and finding the corresponding cutoff of $c\%$, we can calculate the corresponding *TP%*. To illustrate this, if *FP%* is first chosen to be 5%, then 23 non-default firms are misjudged as default ($450 \times 5\% = 22.5$). At the same time, the cutoff $c\%$ is decided. Based on this $c\%$, if we successfully predict four defaulted firms, making $TP\% = 8\% (4/50 = 8\%)$. Thus we

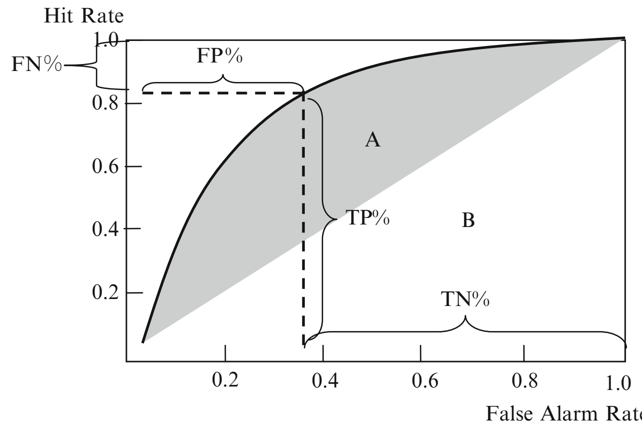

Fig. 62.2 ROC curve

obtain the first set of $(FP\%, TP\%) = (5\%, 8\%)$. Continuing this process, we can get many sets of *FP%* and *TP%*, which generate ROC.

62.3.4 Kolmogorov-Smirnov (KS)

The KS-test tries to determine if two datasets differ significantly. The KS-test has the advantage of making no assumption about the distribution of data. It also enables us to view the data graphically. KS plots the cumulative distribution of default and non-default firms, denoted as $F1$ and $F2$, respective by and then calculates the maximum distance between these two curves as

$$\mathrm{KS} = \max(F1 - F2)$$

The large KS suggest the rejection of the null hypothesis of equality of distributions.

62.3.5 Brier Score

Brier score computes the mean squares error between the estimated and actual default rate. The Brier Score is estimated as

$$B = \frac{\sum_{i=1}^{n} \left(\widehat{P}_i - I_i\right)^2}{n},$$

where $\hat{P}_i$ is the predicted value and I_i is the actual 0 and 1. From the above definition, it follows that the Brier score is always between zero and one. The closer the Brier score is to zero the better is the forecast of default probabilities.

62.4 Source of Data and Empirical Model

62.4.1 Source of Data

To ensure the reliability of financial statements, our samples are actual listed companies on the Taiwan Stock Exchange. Default firms are defined as those stocks that require full delivery; that is, transaction with cash in Taiwan Stock Exchange. These firms include (1) check bouncing of the CEOs, board directors and supervisors of companies; (2) firms that request for financial aid from the government, due to restructuring, bankruptcy, liquidation, ongoing uncertainty, acquisitions, tunneling, trading halts, and credit crunch by banks.

In our sample, there are 52 default companies in the period 1999–2004. For each default company, we search for the three additional companies with a similar size of assets in the same industry, resulting in 156 non-default companies. Hence, 208 companies in total are in our sample. The names and codes of all the companies are reported in Table 62.2 as well as the reasons of their defaults.

We also reserve 20% of our sample for out-of-sample forecast. That is, there are 42 reserved firms.

62.4.2 Empirical Model

Our empirical model is on the basis of Altman's z-score, which contains five variables. Four of them are financial accounting variables and one of is market variable (X_4). Because of multicollinearity, we choose only four of them;[3] that is our model is

$$Y_t = f(X_{1t-1} + X_{2t-1} + X_{3t-1} + X_{4t-1} + 1.0X_{5t-1})$$

where Y is the binary variable with 1 and 0 and 1 denotes defaulted and zero otherwise, X_1 = operation capital/total asset (operating capital = liquid asset − liquid liability), X_3 = earnings before interest and tax (EBIT)/total asset, X_4 = stock value/total liability, X_5 = net sales revenue/total asset (net sales revenues = sales − redemption and discount). All signs are expected to be negative.

62.5 Empirical Results

The left and right parts of Table 62.3 report the estimated results using the logit and the Robust Logit models, respectively. When the logit model is used, all coefficients show the expected negative sign and all are significant except for coefficient of X_5. By contrast, when the Robust Logit model is employed, all coefficients not only show the expected signs but also are significantly different from zero. Alongside this, the pseudo-R^2 is 0.3918 for the logit model but is higher up to 0.9359 for the Robust Logit model, suggesting that in-sample fitting is much better in the Robust Logit model than in the logit model.

Figure 62.3 plots the curve of CAP, ROC, and KS using in-sample forecasts. The curves generated by the Robust Logit method is more concave to the southeast than the logit method shown in the CAP and ROC. With respect to KS method, the maximum distance between non-default and default is also bigger for the Robust Logit method than for the logit method.

Figure 62.4 is similar to Fig. 62.3 but an out-of-sample forecast is used. The CAP and ROC curves generated by the two methods are twisted with each other to some extent and the area under the curves can be hardly evaluated by the human eye. With respect to KS method, the maximum distance of non-defaults and defaults clearly show that the Robust Logit method is superior to the logit method.

Table 62.4 reports the five validation tests by using the in-sample forecast. With respect to the Contingency Table, when the logit method is used, the *TP%* and *TN%* are about 77% but are higher up to 97.67 and 93.67%, respectively, when the Robust Logit method is undertaken. Thus, the Robust Logit method defeats the logit method when the validation is based on the Contingency Table. The KS is 5.288 and 6.410 for the two methods, respectively, again supporting the superiority of the Robust Logit method. The CAP ratio also reaches the similar conclusion, where they are 0.7040 and 0.8308 for the two methods, respectively. Not surprisingly, ROC ratios also support the same conclusion as the two ratios are 0.8447 and 0.9867, respectively. Finally, the Brier score, whose definition is opposite to the previous validation tests, is smaller if the performance of the method is superior. The scores for two models are respectively 0.1207 and 0.0226. Accordingly, all validation tests suggest that the Robust Logit method is superior to the logit method in in-sample prediction.

Table 62.5 reports the validation tests by using the out-of-sample forecast. The superior performance of the Robust Logit method in in-sample forecast becomes less pronounced here. When Contingency Table is employed, the *TP%* and *TN%* yielded by the logit model are about 75%, which is similar to their in-sample counterparts reported in Table 62.4. The values, however, change dramatically when the Robust Logit is used. The *TP%* becomes 100.0% but *TN%* is only about 48%. This implies that the Robust Logit method is aggressive in the sense that it has a greater tendency to assign companies as default. The use of KS test still support the conclusion reached by the in-sample case, i.e., the logit method performs worse than the Robust Logit method. The differences between the two methods in CAP test become trivial as the logit method is 0.6566 and the Robust Logit is

[3] We omit X_2 = retained earnings/total assets.

Table 62.2 All sample companies

Code of failing companies	Names of companies	Default date	Types of default	Matching samples of non-default companies	Data year
9913	MHF	1999/1/18	G	9911(SAKURA), 9915(NienMade), 9914(Merida)	1998
2005	U-Lead	1999/1/24	G	2002(CSC), 2006(Tung Ho Steel), 2007(YH)	1998
2539	SAKURAD	1999/3/22	G	2501(CATHAY RED), 2504(GDC), 2509(CHAINQUI)	1998
2322	GVC	1999/4/1	O	2323(CMC), 2324(Compal), 2325(SPIL)	1998
2522	CCC	1999/4/18	C	2520(KINDOM), 2523(DP), 2524(KTC)	1998
1431	SYT	1999/5/21	H	1432(TAROKO), 1434(F.T.C.), 1435(Chung Fu)	1998
1808	KOBIN	1999/5/24	H	1806(CHAMPION), 1807(ROMA), 1809(China Glaze)	1998
9922	UB	1999/10/5	G	9918(SCNG), 9919(KNH), 9921(Giant)	1998
1206	TP	1999/11/2	H	1216(Uni-President), 1217(AGV), 1218(TAISUN)	1998
1209	EHC	2000/3/23	N	1201(Wei-Chuan), 1203(Ve Wong), 1207(CH)	1999
2528	CROWELL	2000/4/28	G	2526(CEC), 2527(Hung Ching), 2530(DELPHA)	1999
1462	TDC	2000/7/11	G	1458(CHLC), 1459(LAN FA), 1460(EVEREST)	1999
2703	Imperial	2000/9/5	H	2702(HHG), 2704(Ambassador), 2705(Leo Foo)	1999
1422	MICDT	2000/9/6	C	1417(CARNIVAL), 1418(TONG-HWA), 1419(SHINKO.SPIN.)	1999
1505	YIW	2000/9/6	G	1503(Shihlin), 1504(TECO), 1506(Right Way)	1999
2334	KFC	2000/9/6	C	2333(PICVUE), 2335(CWI), 2336(Primax)	1999
2518	EF	2000/9/6	G	2514(LONG BON), 2515(BES), 2516(New Asia)	1999

(continued)

Table 62.2 (continued)

Code of failing companies	Names of companies	Default date	Types of default	Matching samples of non-default companies	Data year
2521	HCC	2000/9/8	G	2505(ky), 2509(CHAINQUI), 2511(PHD)	1999
2019	Kuei Hung	2000/9/16	G	2020(MAYER PIPE), 2022(TYCOONS), 2023(YP)	1999
2011	Ornatube	2000/10/13	C	2012(CHUN YU), 2013(CSSC), 2014(CHUNG HUNG)	1999
9906	Corner	2000/10/27	C	9905(GCM), 9907(Ton Yi), 9908(TGTG)	1999
1222	Yuan Yi	2000/11/2	C	1224(HSAFC), 1225(FOPCO), 1227(QUAKER)	1999
2902	Choung Hsim	2000/11/29	H	2903(FEDS), 2910(CHUN YUAN STEEL), 2912(7-ELEVEN)	1999
2517	CKA-LT	2000/11/30	G	2520(KINDOM), 2523(DP), 2524(KTC)	1999
2537	Ezplace	2001/1/12	G	2534(HSC), 2535(DA CIN), 2536(Hung Poo)	2000
1408	CST	2001/4/2	G	1410(NYDF), 1413(H.C.), 1414(TUNG HO)	2000
1407	Hualon	2001/5/22	G	1402(FETL), 1416(KFIC), 1409(SSFC)	2000
2540	JSCD	2001/5/25	G	2533(YUH CHEN UNITED), 2538(KeeTai), 2530(DELPHA)	2000
2304	A.D.I.	2001/7/28	C	2301(LTC), 2302(RECTRON), 2303(UMC)	2000
1438	YU FOONG	2001/8/10	E	1435(Chung Fu), 1436(FUI), 1437(GTM)	2000
1450	SYFI	2001/8/24	G	1451(NIEN HSING), 1452(HONG YI), 1453(PREMIER)	2000
2318	Megamedia	2001/9/28	G	2315(MIC), 2316(WUS), 2317(HON HAI)	2000
2506	PCC	2001/10/16	G	2514(LONG BON), 2515(BES), 2516(New Asia)	2000

(continued)

Table 62.2 (continued)

Code of failing companies	Names of companies	Default date	Types of default	Matching samples of non-default companies	Data year
1613	Tai-I	2001/10/22	E	1614(SANYO), 1615(DAH SAN), 1616(WECC)	2000
2512	Bao-Chen	2002/4/16	C	2504(GDC), 2509(CHAINQUI), 2511(PHD)	2001
1805	KPT	2002/6/2	G	1802(TG), 1806(CHAMPION), 1807(ROMA)	2001
1602	PEW	2002/9/6	G	1601(TEL), 1603(HwaCom), 1604(SAMPO)	2001
1221	CCI	2003/3/6	C	1217(AGV), 1234(HEYSONG), 1231(Lian Hwa Foods)	2002
2342	MVI	2003/4/18	G	2344(WEC), 2345(ACCTON), 2347(Synnex)	2002
3053	DING ING	2003/4/26	E	3045(TWN), 3046(AOpen), 3052(APEX)	2002
2329	OSE	2003/6/30	G	2330(TSMC), 2331(Elitegroup), 2332(D-LINK)	2002
1212	SJI	2003/9/30	G	1204(jingjing), 1216(Uni-President), 1218(TAISUN)	2002
3001	KIM	2004/3/5	C	3010(WAN LEE), 3011(JH), 3018(TUNG KAI)	2003
2525	Pao Chiang	2004/3/20	G	2520(KINDOM), 2523(DP), 2524(KTC)	2003
2494	Turbocomm	2004/4/15	E	2489(AMTRAN), 2488(HANPIN), 2492(WTC)	2003
2398	Procomp	2004/6/15	H	2382(QCI), 2388(VIA), 2409(AUO)	2003
3021	Cradle	2004/7/26	C	3020(USTC), 3022(ICP), 3023(Sinbon)	2003
2491	Infodisc	2004/8/23	G	2308(DELTA), 2311(ASE), 2312(KINPO)	2003
2490	Summit	2004/9/15	C	2349(RITEK), 2350(USI), 2351(SDI)	2003

(continued)

Table 62.2 (continued)

Code of failing companies	Names of companies	Default date	Types of default	Matching samples of non-default companies	Data year
3004	NAFCO	2004/9/23	H	2356(INVENTEC), 2357(ASUSTEK), 2358(MAG)	2003
1534	Tecnew	2004/9/24	C	1531(SIRUBA), 1532(CMP), 1533(ME)	2003
9936	Compex	2004/10/20	E	9933(CTCI), 9934(GUIC), 9935(Ching Feng)	2003

The number is the code listed in Taiwan Stock Exchange
Type of Default: *C* check bounce; *E* concern of continuing operation; *G* financial Aid from government; *O* substantial loss (low net worth)

Table 62.3 Estimated results: logit vs. Robust Logit

Methods	Logit Coefficients	t-value	Robust Logit Coefficients	t-value
Constant	−0.3600	−0.7506	17.0487**	2.1627
X_1	−1.6195	−1.1766	−10.2357*	−1.7913
X_3	−13.1535***	−4.1651	−234.2707**	−2.2311
X_4	−0.5519**	−2.0683	−2.1146**	−2.3319
X_5	−0.4227	−0.5858	−12.3312**	−2.0225
Log likelihood	−61.4865		−9.510	
Average likelihood	0.6905		0.9200	
Pseudo-R-square	0.3918		0.9359	
Number of sample	166		114	
Number of default companies	43		43	
Number of non-default companies	123		71	
Medium of residuals	–		6.42811e-04	

*,**,*** denote significant at 10, 5 and 1% level, respectively

only 0.6812. Similar results occur in ROC test as the former test is 0.6717 but the latter one is 0.6883. Thus, based on CAP and ROC, they are in a tie. Last, to our surprise, the Robust Logit method is defeated by the logit method as its Brier score is higher than the logit method. Thus, when the out-of-sample forecast is implemented, the differences between two methods are hard to distinguish.

62.6 Conclusion

We compare the forecast ability between logit and Robust Logit methods, where the latter take the possible outliers into account. Six validation tests are employed when the in-sample forecasts are compared, i.e., pseudo-R square, Contingency Table, CAP, ROC, KS and Brier score, whereas the latter five validation tests are undertaken for the out-of-sample forecast.

With respect to the in-sample forecasts, Robust Logit method is substantially superior to the logit method when using all validation tests here. With respect to the out-of-sample forecasts, Robust Logit method yields less type II but large type I errors than the logit method when Contingency Table is used, suggesting that Robust Logit is more aggressive in assigning firms as default. Robust Logit is marginally better than the logit method when CAP, ROC, and KS, are adopted but worse when the Brier score is used. Thus, the superiority of Robust Logit is less pronounced or even disappears in the out-of-sample forecasts.

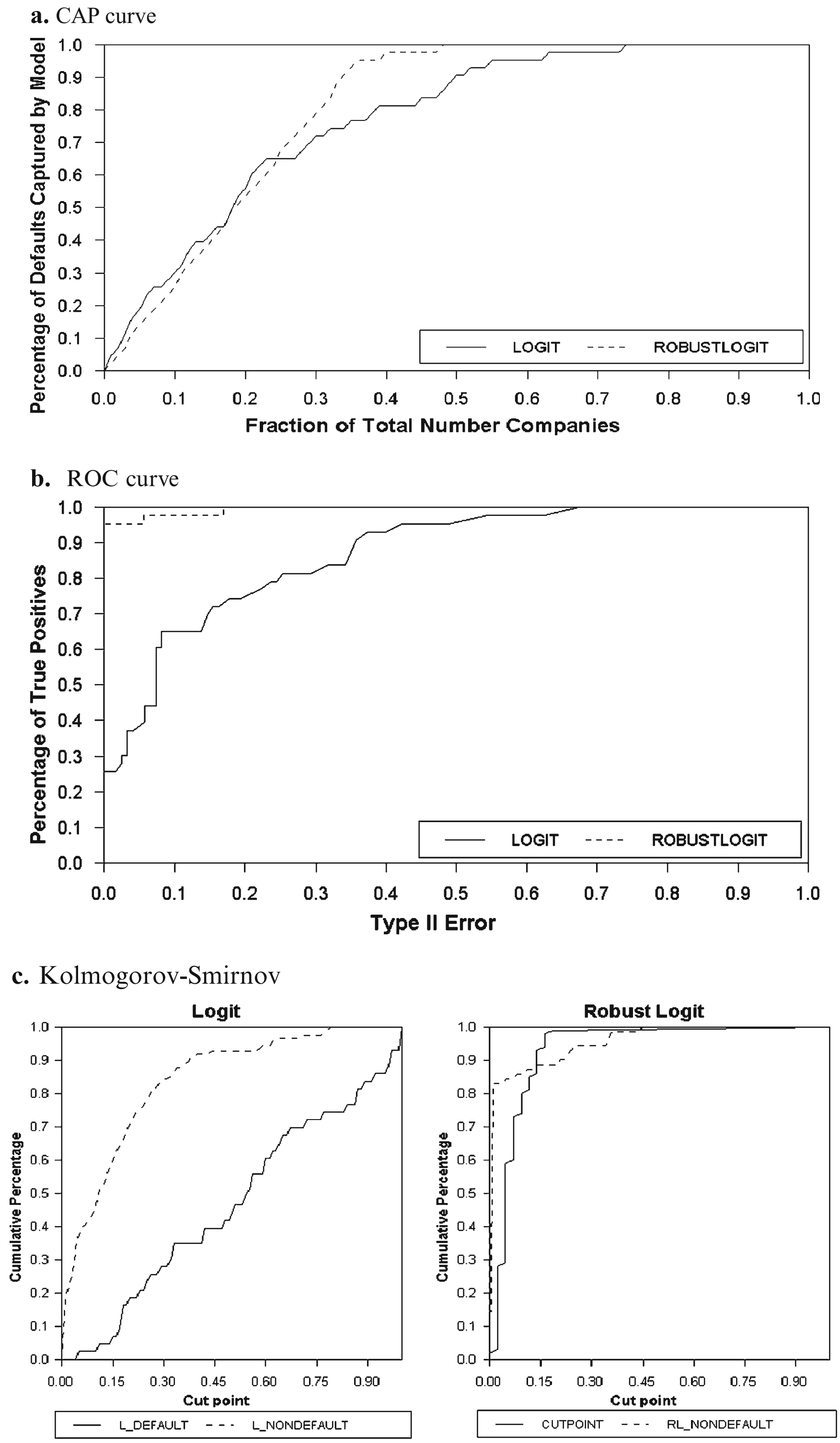

Fig. 62.3 In-sample forecast: Logit and Robust Logit

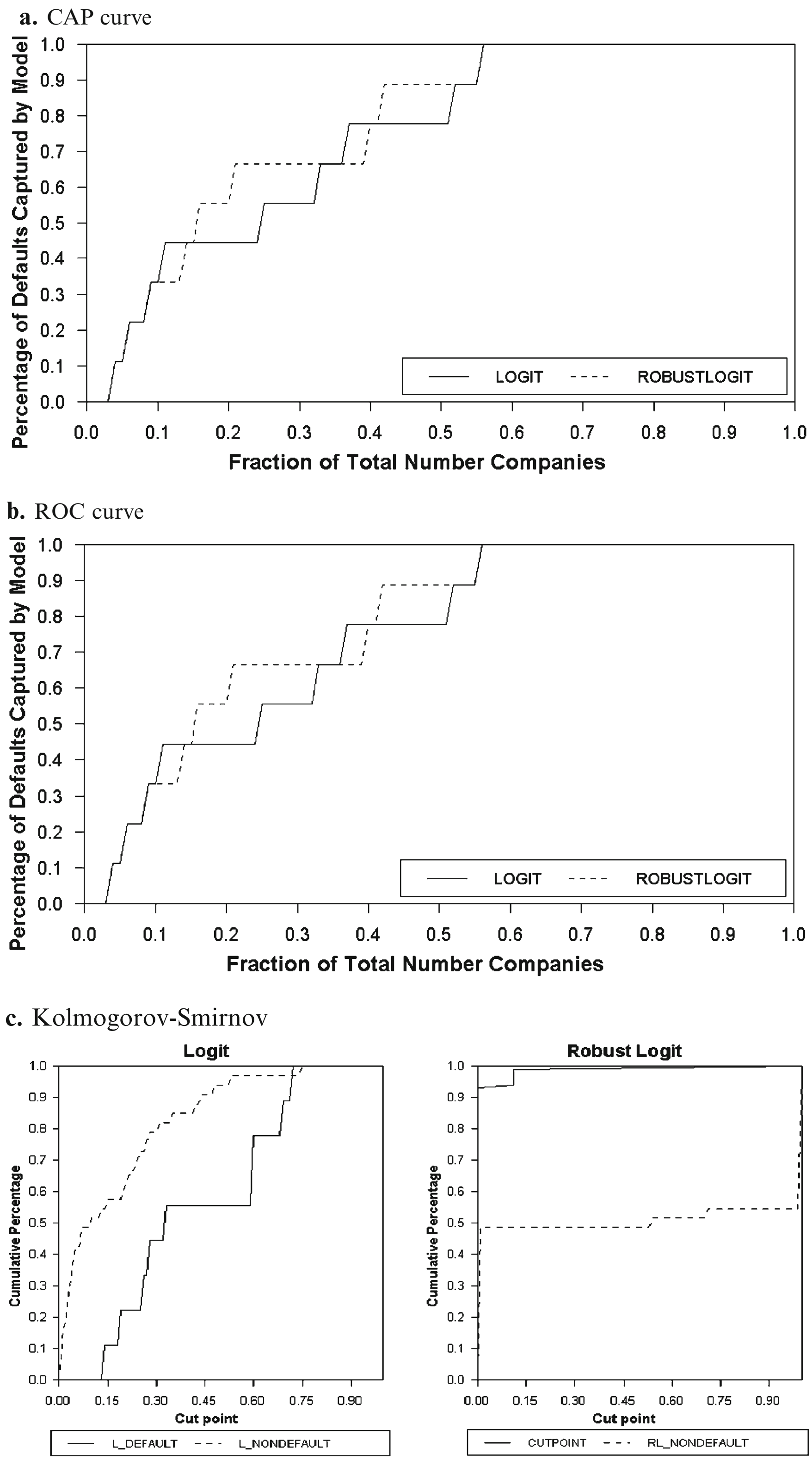

Fig. 62.4 Out-of-sample forecast: logit and Robust Logit forecast

Table 62.4 Validation tests (in-sample forecast)

Methods	Logit method		Robust Logit method	
Cross-classification	TP% 76.74%	TN% 78.05%	TP% 97.67%	TN% 94.37%
KS	5.288		6.410	
CAP	0.7040		0.8308	
ROC	0.8447		0.9867	
Brier score	0.1207		0.0226	

CAP: cumulative accuracy profile; *ROC*: receiver operating curve; *KS*: Kolmogorov-Smirnov

Table 62.5 Estimated results of cross-classification, KS and Brier (out-of-sample forecast)

Methods	Logit method		Robust Logit method	
Cross-classification	TP% 77.78%	TN% 72.73%	TP% 100.00%	TN% 48.48%
KS	1.45		1.558	
CAP	0.6566		0.6812	
ROC	0.6717		0.6883	
Brier score	0.1319		0.3756	

CAP: cumulative accuracy profile; *ROC*: receiver operating curve; *KS*: Kolmogorov-Smirnov

References

Altman, E. I. 1968 "Financial ratios, discriminate analysis and the prediction of corporate bankruptcy." *Journal of Finance* 23(4), 589–609.

Altman, E. I. and A. Saunders 1998 "Credit risk measurement: developments over the last 20 years." *Journal of Banking & Finance* 21(11–12) 1721–1742.

Atiya, A. 2001 "Bankruptcy prediction for credit risk using neural networks: a survey and new results." *IEEE Transactions on Neural Networks* 12(4), 929–935.

Atkinson, A. C. 1985 *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*, Oxford University Press, New York.

Atkinson, A. C. 1994 "Fast very robust methods for the detection of multiple outliers." *Journal of the American Statistical Association* 89(428), 1329–1339.

Atkinson, A. C. and T. -C. Cheng 2000. "On robust linear regression with incomplete data." *Computational Statistics & Data Analysis* 33(4), 361–380.

Atkinson, A. C. and M. Riani 2001 "Regression diagnostics for binomial data from the forward search." *The Statistician* 50(1), 63–78.

Atkinson, A. C. and M. Riani 2006 "Distribution theory and simulations for tests of outliers in regression." *Journal of Computational Graphical Statistics* 15(2) 460–476.

Barrett, B. E. and J. B. Gray 1997 "Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression." *Computational Statistics and Data Analysis* 26(1) 39–52.

Cook, R. D. and S. Weisberg. 1982. *Residuals and influence in regression*, Chapman & Hall/CRC, London.

Dimitras, A. I., S. H. Zanakis and C. Zopoundis 1996 "A survey of business failures with an emphasis on prediction methods and industrial application." *European Journal of Operation Research* 90(3), 487–513.

Doumpos, M. and C. Zopounidis 2002. "Multi-criteria classification and sorting methods: a literature review." *European Journal of Operational Research* 138(2), 229–246.

Doumpos, M., K. Kosmidou, G. Baourakis and C. Zopounidis 2002 "Credit risk assessment using a multicriteria hierarchical discrimination approach: a comparative analysis." *European Journal of Operational Research* 138(2), 392–412.

Flores, E and J. Garrido 2001 "Robust logistic regression for insurance risk classification." *Business Economics* Universidad Carlos III, Departamento de Economía de la Empresa Working Papers wb016413.

Haslett, J. 1999. "A simple derivation of deletion diagnostic results for the general linear model with correlated errors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(3), 603–609.

Lennox, C 1999 "Identifying failing companies: a reevaluation of the logit, probit and DA approaches." *Journal of Economics and Business* 51(4), 347–364.

Levine, R. and S. Zervos 1998 "Stock markets, banks, and growth." *American Economic Review* 88(3), 537–558.

Ohlson J. T. 1980 "Financial ratios and the probabilistic prediction of bankruptcy." *Journal of Accounting Research* 18(1), 109–131.

Piramuthu, S. 1999 "Financial credit-risk evaluation with neural and neurofuzzy systems." *European Journal of Operational Research* 112(2) 310–321.

Riani, M. and A. Atkinson 2007 "Fast calibrations of the forward search for testing multiple outliers in regression." *Advances in Data Analysis and Classification* 1(2) 123–141.

Rousseeuw P. J. 1983 "Regression techniques with high breakdown point." *The Institute of Mathematical Statistics Bulletin* 12, 155.

Rousseeuw, P. J. 1984 "Least median of squares regression." *Journal of the American Statistical Association* 79(388), 871–880.

Rousseeuw, P. J. and A. Christmann 2003 "Robustness against separation and outliers in logistic regression." *Computational Statistics & Data Analysis* 43(3) 315–332.

Rousseeuw P. J. and V. J. Yohai 1984 "Robust regression by means of S-estimators," in *Robust and nonlinear time series analysis*, Vol. 26, W. H. Franke and R. D. Martin (Eds.). Springer, New York, pp. 256–272.

Sobehart, J. R. and S. C. Keenan 2004 "Performance evaluation for credit spread and default risk models," in *Credit risk: models and* *management* Second Edition, D. Shimko (Ed.). Risk Books London, pp. 275–305.

Sobehart, J. R., S. C. Keenan and R. M. Stein 2000 "Rating methodology: benchmarking quantitative default risk models: a validation methodology." *Moody's Investors Service, Global Credit Research*, New York.

Stein, R. M. 2002 "Benchmarking default prediction models: pitfall and remedies in model validation." *Moody's KMV, Technical Report*, No. 030124.

Westgaard, S and N. Wijst 2001 "Default probabilities in a corporate bank portfolio: a logistic model approach." *European Journal of Operational Research*, 135(2), 338–349.

Wu, C. and X. M. Wang 2000 "A neural network approach for analyzing small business lending decisions." *Review of Quantitative Finance and Accounting*, 15(3), 259–276.

Zhu A. M. Ash and R. Pollin 2002 "Stock market liquidity and economic growth: a critical appraisal of the Levine/Zervos model." *International Review of Applied Economics* 18(1), 1–8.