Statistical Credit Rating Methods

Prof. Jin-Chuan Duan
Risk Management Institute
& NUS Business School
rmidjc@nus.edu.sg

Prof. Keshab Shrestha
Risk Management Institute
rmikms@nus.edu.sg

INTRODUCTION

Default is one of the most important events that can take place in the life of a firm. It has grave consequences for all stakeholders, from shareholders and debt-holders to suppliers and employees. The defaults of high profile firms such as Enron, World-Com and Lehman Brothers, clearly demonstrate how catastrophic such an event can be. The extent of the damage caused to both economy and society will of course depend on the size and the systemic nature of the defaulting firm. Nevertheless, even for a small firm without any systemic implications, default will be an event of great importance to the stakeholders.

Assessing default risk is obviously an important step towards understanding the default risk of an obligor and/or a portfolio of obligors. An in-depth knowledge of credit risk helps users devise tools to mitigate the negative impacts of a default. However, analysts have differing views on how credit analysis should be best conducted. Some use the so-called fundamental approach, which relies on dissecting various accounting and operating ratios as well as devising some qualitative measures to reflect judgments. Rules of thumb are then established, based on their experiences. Others contend that credit analysis can be approached scientifically; for example, statistical tools can be applied to a sample of past defaults, thus attempting to find a relationship between the defaults and the attributes of obligors. The conceptual difference between the two approaches may be reduced to the classic dichotomy of art and science. It is our contention, however, that statistical tools are indispensable to default analysis. Examining the evidence, their power to assess corporate default likelihood is indisputable. The statistically-based quantitative model typically requires more effort in the development stage, but the operational costs are lower simply because of its lower manpower requirements. By an intelligent application of the statistical credit rating method, the quality of credit assessment can be greatly improved at comparatively low additional costs.

Measurement of default risk or estimation of default probability involves studying the tail of a probability distribution. From a statistical point of view, estimating the tail probability is, unsurprisingly, difficult. The difficulties arise from the fact that rare events are rarely observed. To obtain a sample with a number of default cases large enough for a meaningful statistical analysis, one must gather a very large data sample, which means observing many firms over a comparatively long time period.

The default of a firm is likely to be influenced by the state of the economy; for example, we would expect more defaults during recessionary times than during expansionary times. We would also naturally assume a firm's vulnerability to default to be related to the state of its financial health. If a firm has higher liquidity, for example, it is more likely to meet its debt service needs. Likewise, a firm which makes little use of debt to finance its business operations is unlikely to face any serious debt service pressure and hence runs little risk of default. To quantify default risk, one must have access to observations of risk factors that are common to all firms in the economy, in order to discern behavior characteristic of defaulters, over time. In addition, one must rely on the known attributes of individual firms to characterize defaults across firms. In this paper, we assume that a reasonably large sample of firms can be obtained, and that the data sample forms a panel of cross-sectional and time series observations containing common risk factors and individual firms' attributes over a period of time. The data sample is assumed to be sufficiently rich to statistically model default behavior.

We discuss some of the well-known statistical credit rating methods available for practitioners to use. In addition, we introduce some of the more recent advances in default prediction methodology. The methods discussed include the traditional Z-score method proposed by Altman (1968), the logistic regression, the artificial neural network and the support vector machine, and the Poisson intensity model. We use a sample of Japanese firms obtained from the credit rating database of the Risk Management Institute (RMI) of the National University of Singapore to demonstrate these credit rating methods. Our demonstration dataset starts from 1997 and ends in 2009 inclusive. The dataset comprises 45,108 firm-year observations with 165 default cases. However, depending on the model used in the estimation, the sample size may vary. The actual sample size also depends on the number of missing values for the variables used in a model.

We assess a model's performance using the cumulative accuracy profile (CAP) and the receiver operating characteristic (ROC). These two popular ways of assessing a model's performance can be applied to any method that can produce rank orders. Naturally, all credit rating methods are expected to rank obligors. The CAP and ROC are in effect two equivalent performance metrics which yield the same performance conclusion. We will discuss these performance metrics and apply them to assessing credit rating models.

I. ALTMAN'S Z-SCORE AND PERFORMANCE METRICS

Business communities have been interested in credit risk analysis ever since debt capital began to be used to finance business undertakings. The first organized credit rating service may have been the Mercantile Agency, forerunner of Dun and Bradstreet, which was established on July 20, 1841 by an enterprising businessman named Lewis Tappan. One of the goals of the Mercantile Agency was to provide reliable, consistent, and objective credit information by forming a network of correspondents.[1]

Most of the early studies of bankruptcy used some accounting ratios whereby the accounting ratios of failed firms were compared to the corresponding ratios of the non-failed firms in the sample. Some of these studies analyzed the time series trends of these ratios of failed to non-failed firm, before they failed. As pointed out by Beaver (1966), these studies include Fitz Patrick (1932) and Winakor and Smith (1935). In another study, Merwin (1942) compared three financial ratios of continuing firms with those of discontinued firms for the period from 1926 to 1936. The financial ratios in Merwin's study include current asset to current liabilities, net worth to total debt and net working capital to total assets. Merwin (1942) found that the difference in means was noticeable as early as six years before the discontinuance, and the difference increased as the year of discontinuance approached.

Beaver (1966) used financial ratios in an extensive manner to predict bankruptcy. He used different threshold points for different ratios to minimize the percentage of incorrect predictions. He also considered the Type I error (misclassification of a failed firm) and Type II error (misclassifying a non-failed firm) in the analysis. He also used the likelihood ratios in the analysis. He found, based on the percentage of firms misclassified, that cash-flow to total debt seemed to offer the most accurate prediction, followed by the net income to total assets ratio, when considering a single ratio at a time.

Both these studies used different ratios in isolation rather than jointly. Beaver (1966) actually suggested the use of multi-ratio approach for future research. It is hardly surprising that defaults are too complex to be explained by a single factor. Altman (1968) was the first to use a multi-ratio approach. All methods considered in this paper naturally consider multiple ratios. We will begin with Altman's Z-score.

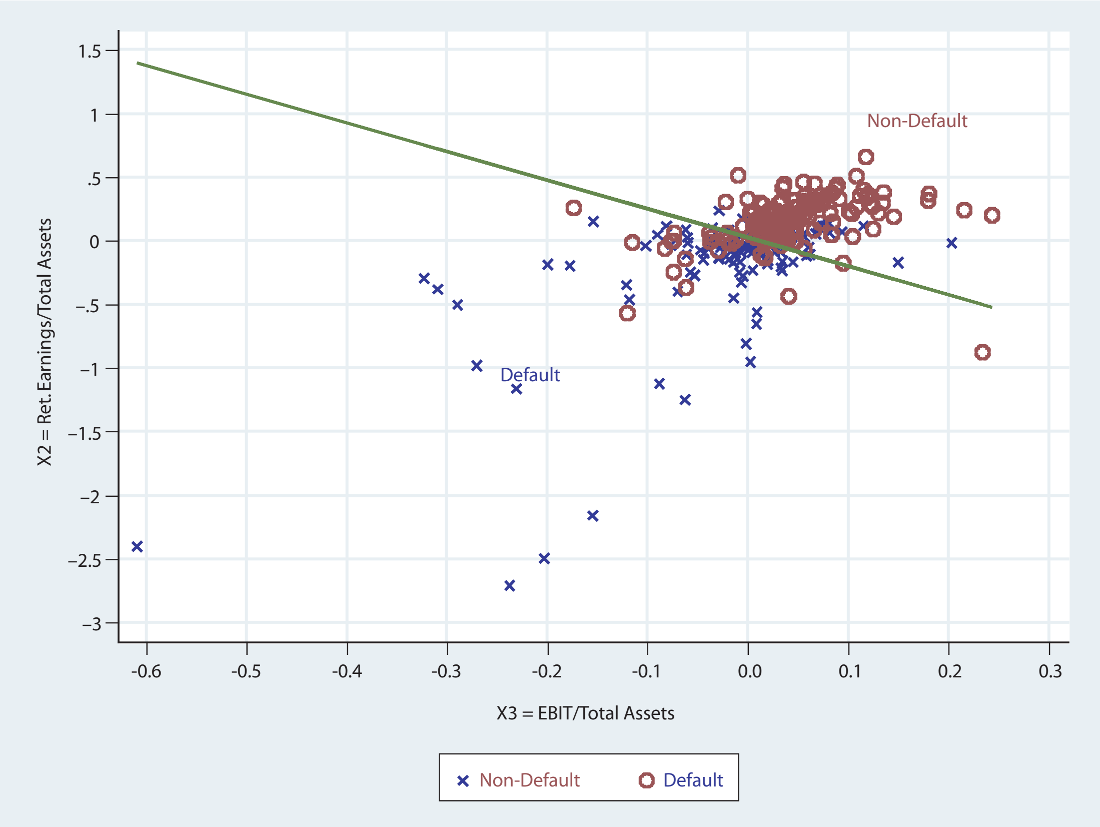
1.1 Altman's Z-score

Altman (1968) is generally regarded as the first researcher to use a statistical model to connect default to different accounting ratios taken together instead of one ratio at a time. Specifically, he employed a linear multiple discriminant analysis (MDA) based on five accounting ratios to classify firms that went bankrupt separately from firms that did not.

The MDA can be understood by using a simple example where only two firm characteristics – Retained earnings/Total asset and EBIT/Total assets – are used to classify firms into default and non-default groups. Figure 1 shows the distribution of these two accounting ratios for firms that defaulted and firms that did not default. The MDA finds a linear function of these two variables, known as the discriminant function, so that most of the firms that defaulted will be on one side of the discriminant function whereas most of the firms that did not default will be on the other side. In Figure 1,

FIGURE 1

Distribution of Retained Earnings/Total Assets and EBIT/Total Assets for firms that defaulted and those that did not



the discriminant function is represented by the straight line that best separates the two groups.

The classification based on the MDA would be perfect if all defaulted firms had been on one side of the discriminant function and all surviving firms had been on the other side. In such a situation, there would be no misclassification for the in-sample analysis. In practice, however, it is generally not possible to find variables that give rise to a perfect discriminant function, separating completely the defaulted firms from those which did not default. The best realistic scenario one should hope for is to find variables that generate some minimum overlap. As seen in Figure 1, there are some firms of each type, bankrupt and non-bankrupt, sitting on the wrong side of the discriminant function.

We now describe the Z-score method proposed by Altman (1968). Suppose there are *m* accounting ratios or discriminating variables, X_1, X_2, …, X_m, and we want to use them to come up with a credit rating model. Altman's Z-score method uses the discriminating variables in a linear fashion. The analysis involves estimation of the following discriminant function:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m \qquad (1)$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ are the unknown parameters. In order to formally discuss the MDA, we need to introduce the following notations:

X_{ikj} = value of the ith accounting ratios for firm *j* in group *k*
N_k = number of firms in group *k*
Z_{kj} = value of the discriminant function for firm *j* in group *k,* i.e.,

$$Z_{kj} = \beta_0 + \beta_1 X_{1kj} + \beta_2 X_{2kj} + \ . \ . \ . \ + \beta_m X_{mkj} \qquad (2).$$

Note that there are only two groups: default and non-default. One can consider two types of variation in scores. The within-group variation can be captured by the sum-of-squares S_w whereas the between-group one is reflected in the sum-of-squares S_B. They are respectively defined as follows:

$$S_W = \sum_{k=1}^{2}\sum_{j=1}^{N_k}\left(Z_{kj} - \bar{Z}_k\right)^2, \quad \bar{Z}_k = \frac{1}{N_k}\sum_{j=1}^{N_k} Z_{kj} \qquad (3)$$

$$S_B = \sum_{k=1}^{2} N_k\left(\bar{Z}_k - \bar{\bar{Z}}\right)^2, \quad \bar{\bar{Z}} = \frac{1}{N_1 + N_2}\sum_{k=1}^{2}\sum_{j=1}^{N_k} Z_{kj} \qquad (4).$$

Note that $\bar{Z}_k$ is the sample mean of the scores for group *k* with the sample size N_k and $\bar{\bar{Z}}$ is the grand sample mean of the scores for the whole sample with the sample size equal to $N = N_1 + N_2$. The scores can be computed only when the parameter values are available. The unknown parameters $\beta_1, \beta_2, \ . \ . \ . \ , \beta_m$ have to be estimated based on some sensible criterion. Altman (1968) used an *F*-statistic type criterion, which maximizes the ratio of the between-group sum-of-squares to the within-group sum-of-squares:[2]

$$\lambda = \frac{S_B}{S_W} \qquad (5).$$

Since the ratio of the sums-of-squares in equation (5) is independent of the first unknown parameter β_0, this parameter cannot be determined by maximization. It can be, for example, chosen so that the average score for the whole sample equals zero. In the case of Altman's Z-score construction, β_0 is actually set to zero instead of making the average score zero. It is also clear that the parameter values can only be determined up to a scalar multiplier. The choice of the scalar does not affect the result, but may make the resulting scores easier to comprehend. For example, one can make a safer firm to have a higher score or turn the rank order around by simply multiplying all scores by a negative constant. Altman (1968) chose the scalar in such way that a safer obligor will have a better score.

It is worth noting that using an *F*-statistic criterion as in equation (5) is in effect the same as the commonly used Fisher's linear discriminant analysis. The target ratio in Fisher's linear discriminant analysis may appear to differ from that of equation (5). It actually only differs by a multiplicative constant after some algebraic manipulation. Therefore, the optimal solutions for the parameter values are the same.

Out of 165 defaults in our Japan data sample, 143 cases come with the complete accounting data that are needed for Altman's Z-score model. Following Altman (1968), we obtain the same number of firms that did not default by matching the industry and year of default. In other words, we use some kind of paired sample design. Therefore, our total sample consists of 286 firms, half of which belong to the default group and half of which belong to the non-default group. This sample will be referred to as the matched sample.

As in Altman (1968) we use the following five accounting ratios as the discriminating variables:

X_1 = Working capital/Total assets
X_2 = Retained earnings/Total assets
X_3 = EBIT/Total assets
X_4 = Market value of equity/Book value of the total Debt
X_5 = Sales/Total assets

The accounting data for the defaulted firms are taken from the annual financial statement available prior to the default date. The summary statistics for our sample are given in Table 1. A separate set of summary statistics for the default and non-default firms are given in Table 2. It is clear from Table 2 that the sample means and medians for all five variables are higher for the non-default firms as compared to the default firms. Therefore, it is quite reasonable to expect that these five variables are informative about defaults.

Applying the MDA on our matched sample leads to the following discriminant function[3]:

$$Z = -1.195 + 0.820X_1 + 1.128X_2 + 3.338X_3 + 0.406X_4 + 0.767X_5 \quad (6).$$

As expected, all five discriminating variables have positive coefficients implying that higher values of these variables are likely to be associated with firms in the non-bankrupt group.

TABLE 1
Summary statistics of the five discriminating variables for the whole sample that consists of 143 default firms and 143 non-default firms

	N	Mean	Std. Dev.	Median
Working capital/Total assets	286	0.0523	0.322	0.103
Retained Earnings/Total assets	286	0.0125	0.394	0.057
EBIT/Total assets	286	0.0194	0.084	0.028
Market value equity/Book value of total Debt	286	0.9568	2.753	0.245
Sales/Total assets	286	1.0231	0.560	0.856

TABLE 2
Summary statistics of the five discriminating variables for firms in either the default or non-default group

	N	Mean	Std. Dev.	Median
Default Group				
Working capital/Total assets	143	–0.0928	0.363	–0.051
Retained Earnings/Total assets	143	–0.1638	0.463	–0.038
EBIT/Total assets	143	–0.0162	0.095	0.004
Market value equity/Book value of total Debt	143	0.3004	0.828	0.123
Sales/Total assets	143	0.9331	0.564	0.838
Non-Default Group				
Working capital/Total assets	143	0.197	0.186	0.197
Retained Earnings/Total assets	143	0.189	0.187	0.164
EBIT/Total assets	143	0.055	0.052	0.050
Market value equity/Book value of total Debt	143	1.613	3.695	0.520
Sales/Total assets	143	1.113	0.543	0.891

The effectiveness of a classification method, or a credit rating method in the present context, can be measured either by the cumulative accuracy profile or by the receiver operating characteristics, both of which are used in this paper.[4] We will discuss the concepts underlying these two measures and apply them to Altman's Z-score first.

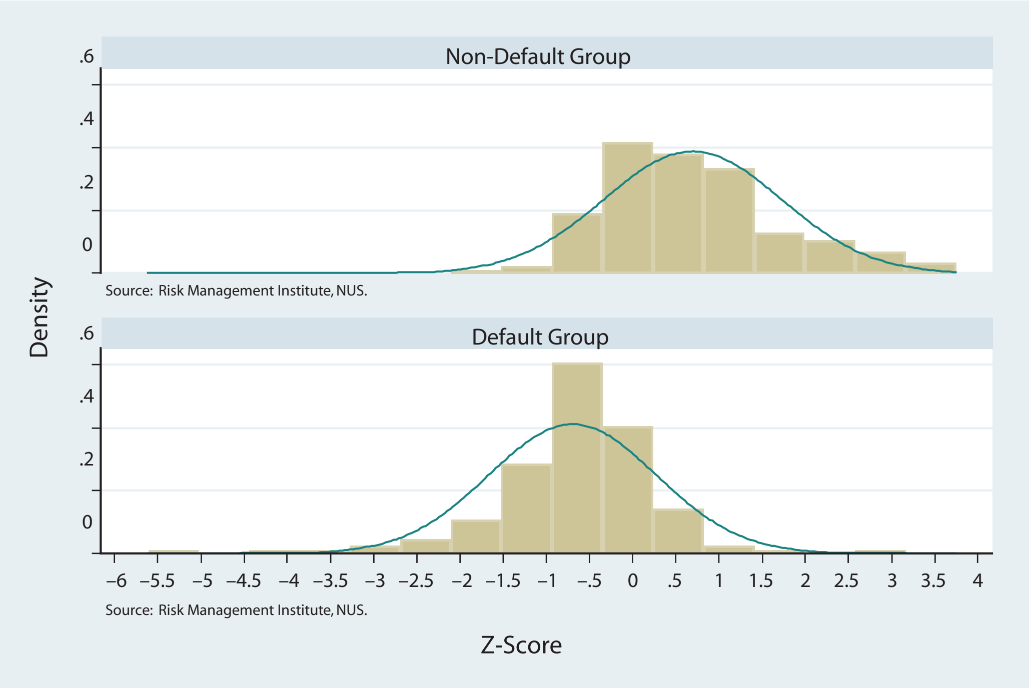
1.2 Cumulative Accuracy Profile (CAP)

The CAP is obtained by first ordering the Z-scores from the lowest to highest values where a lower value means a higher probability of default. Then, for a given fraction *x* of the total number of firms, the CAP curve indicates the fraction of the defaulted firms whose Z-scores are less than or equal to the maximum Z-score up to fraction *x*, where fraction *x* will be varied from 0% to 100%. In our sample, for example, the 50th percentile value of the Z-score is –0.0784; that is, 143 out of 286 firms have their Z-scores less than or equal to –0.0784. Out of these 143 firms, 114 firms turned out to be the defaulted ones, which then constitutes approximately 79.72% (114 out of 143) of the defaulted firms. Therefore, the value of CAP at 50% equals 79.72%.

The CAP associated with Altman's Z-score is shown in Figure 3 where the CAP associated with a random model is also displayed. A random model obviously has no ability to distinguish the firms in the non-default group from those in the default group, and as a result, its CAP should be a straight line with a 45 degree angle. The CAP of a perfect model is also given in Figure 3, which rises quickly to 100% once the number of firms under consideration reaches the number of defaulted firms. The effectiveness of a rating model can be measured by the area between the CAP of the rating model and the CAP of the random model. Let a_R represent the area between the CAP of the rating model being evaluated and the CAP of the random model. Also, let a_P be the area between the CAP of the perfect model and the CAP of the random model. Then, the accuracy ratio (*AR*) is defined as

$$AR = \frac{a_R}{a_p} \qquad (7).$$

FIGURE 2
Histograms of the Z-scores for the default and non-default groups

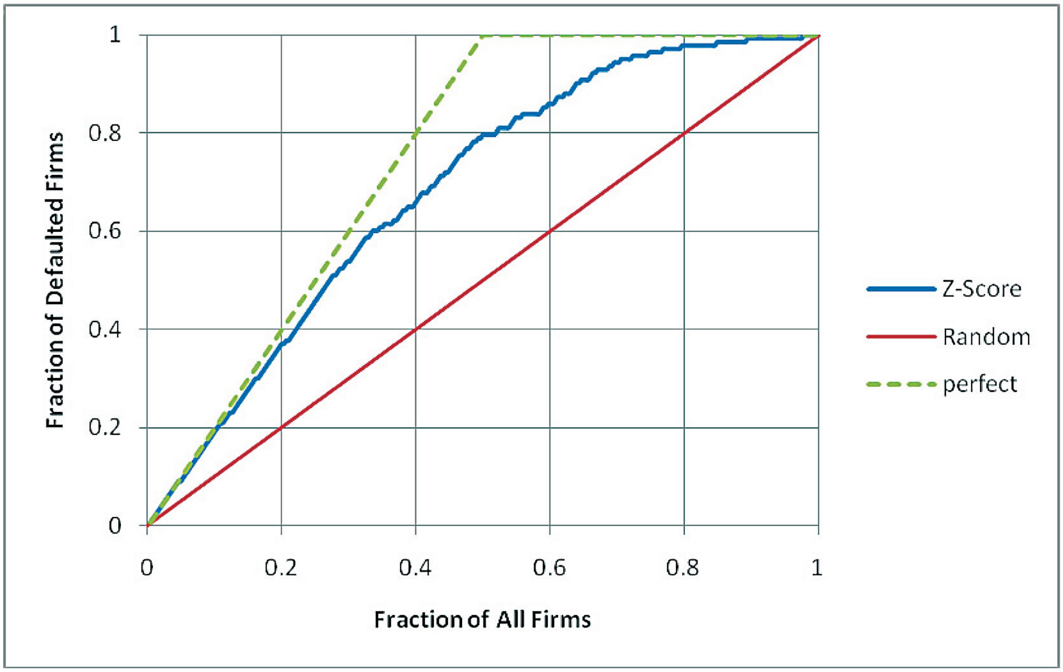


Source: Risk Management Institute, NUS.

Source: Risk Management Institute, NUS.

FIGURE 3

The CAP for the matched sample of 286 firms



Thus, the better a specific credit rating model, the closer the value of its accuracy ratio is to one. For our matched sample of the Japanese firms, the accuracy ratio is approximately equal to

$$AR = \frac{a_R}{a_P} = \frac{0.1837}{0.2500} = 0.7350$$

The CAP and AR computed above are based on the firms in the default group matched by the same number of firms from the non-default group. Instead of taking only the firms from the matched sample, it will be better if we measure the performance of Alman's Z-score method by considering all firm-years in our sample. In total, there are 38,130 firm-year observations for which the discriminating accounting variables are available. We re-compute the CAP and AR for this sample. The resulting CAP is shown in Figure 4. The accuracy ratio for this sample of all firm-year observations is equal to

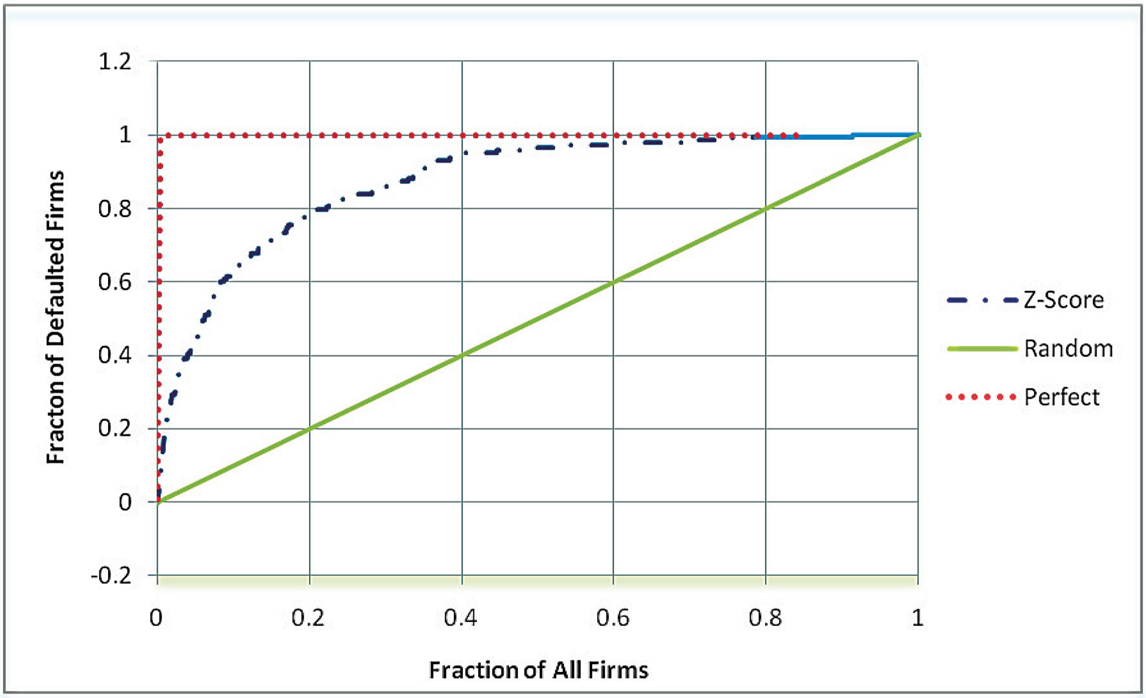
$$AR = \frac{a_R}{a_P} = \frac{0.3745}{0.4981} = 0.7520$$

It is interesting to note that the *AR* is marginally higher for the whole sample as compared to that for the matched sample.

1.3 Receiver Operating Characteristic (ROC)

The use of the ROC has a long history specifically in signal detection theory. In the credit rating context, Sobehart and Keenan (2001) employed the ROC in the evaluation of rating models. When one applies a specific rating model, there will be a rating score for each firm. Under our current consideration, it will be an Altman's Z-score. Once the rating scores are in place, one can classify each firm into the default or non-default group based on some arbitrary cut-off value for the score. For example, at some cut-off value C, we can identify all firms whose scores are less than or equal to C and send them to the default group, and put the remaining firms in the non-default group. Unless the rating model is perfect, it will generate some misclassifications. In other words, some of the defaulted firms may be erroneously classified into the non-default group whereas some surviving firms are

FIGURE 4
The CAP for the whole sample of 38,130 firm-year observations



put into the default group. The ROC is obtained by considering the proportion of correct classifications vs. misclassifications for all possible cut-off values. Four different possible outcomes are shown in Table 3.

There are two possible misclassifications. For example, the firms in the default group may be misclassified as belonging to the non-default group (Miss). This type of error is referred to as Type I error. The second possible misclassification is classifying non-default firms in the default group (False Alarm). This type of error is referred to as Type II error. A classification method is considered superior if it leads to a lower Type I error for any level of Type II error, or vice versa.

TABLE 3
Possible outcomes in classification using cutoff value C

	Classification Decision	
True State	**Non-default**	**Default**
Non- default	Z-score > C (**Correct prediction**)	Z-score ≤ C (**False Alarm:** Type II error)
Default	Z-score > C (**Miss**: Type I error)	Z-score ≤ C (**Hit**)

It is important to note that the Type I and Type II errors depend on cut-off value C. By changing the value of C, we can definitely reduce the Type I error, but it will usually lead to a higher Type II error. Since the Type I and Type II errors depend on the choice of the cutoff value, one should view a suitable performance measure as an overall assessment of the trade-off between the Type I and II errors for the entire range of cutoff values. The ROC is one such measure. Let us define a few notations that are needed for describing the ROC. The first one is the hit rate, HR(C), which naturally depends on cut-off value C. For a given cut-off value C, the hit rate is defined as

$$HR(C) = \frac{H(C)}{N_B} \tag{8}$$

where H(C) is the number of defaulted firms whose scores are less than or equal to C and N_B is the total number of firms in the default group. Therefore, the hit rate is equal to the fraction of the defaulted firms that have been classified correctly using cut-off value C. The next term needed for the ROC is the false alarm rate, FAR(C), which also depends on cut-off value C. The false alarm rate is defined as

$$FAR(C) = \frac{F(C)}{N_{NB}} \tag{9}$$

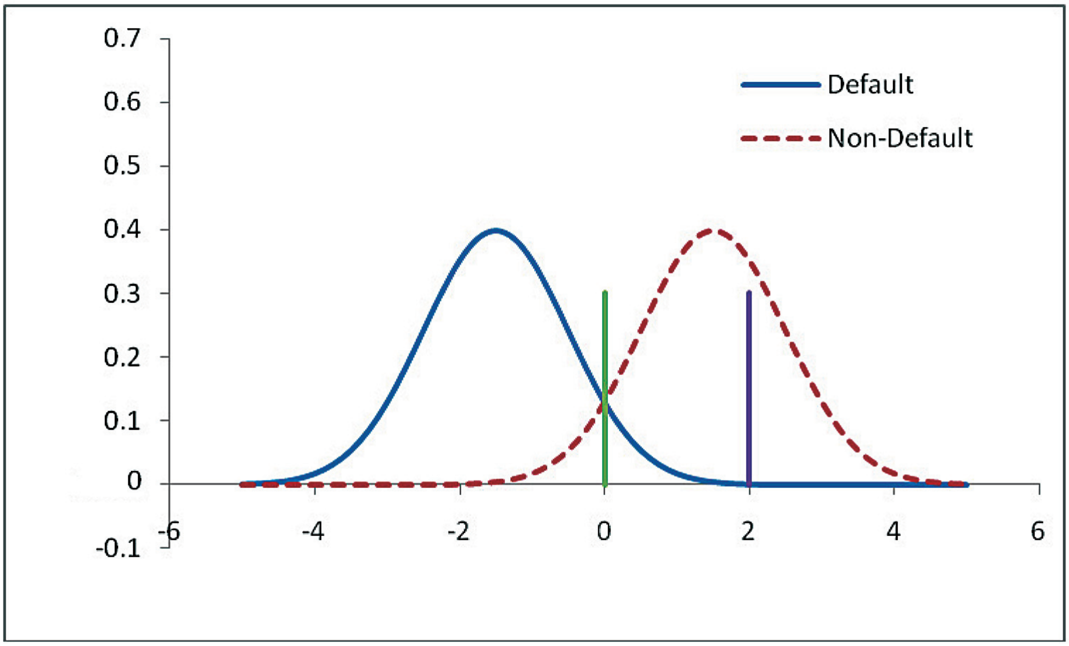
where N_{NB} is the total number of firms in the sample which actually belongs to the non-default group and F(C) is the number of non-default firms that are incorrectly classified as defaulted, i.e., the number of non-default firms whose scores are less than or equal to C.

Figure 5 displays two hypothetical score distributions: one for the default group (solid line) and one for the non-default group (dashed line). For any given cutoff value C, the area to the left of C under the score distribution for the default group is equal to the hit rate, HR(C). Similarly, the area to the left of C under the score distribution for the non-default group is the false alarm rate, FAR(C). For example, if we set the cutoff value to 2.0, the hit rate is almost 100% or the Type I error is zero, i.e. there is no misclassification for the firms in the default group. However, for this cutoff value, the false alarm rate will be very high (more than 50%). Similarly, if we set the cutoff value at 0, then the hit rate will be much lower than 100%, but the false alarm rate will also be reduced significantly. If there is no overlap of the two score distributions, one will be able to find some cutoff value at which the hit rate becomes 100% and the false alarm rate remains at 0.

It is important to note the role played by the cut-off value for the rating scores. If we increase C, then the hit rate, HR(C), will increase. But it also increases the false alarm rate, FAR(C). We can compute the HR(C) and FAR(C) by varying C from the lowest value of the obtained scores to the highest value. The ROC curve is simply a plot of HR(C) vs. FAR(C).

The ROC has some interesting features. First, its relation with the hit rate and the false alarm rate makes it quite intuitively appealing. Second, the area under the ROC curve is asymptotically normally distributed (Bamber, 1975; DeLong, DeLong, and Clarke-Pearson, 1988). Thus, it can be a convenient device for assessing alternative rating models in light of statistical confidence

Figure 5
Distribution of rating scores for the default and non-default groups



intervals. For example, the random model, which has no classification power, has the area under the ROC curve equal to 0.5. Thus, a particular rating model can be tested against the random model as a benchmark.

Table 4 presents the classification table for our Japanese sample using the cutoff value of 0. 18.18% of the defaulted firms have been classified as non-default firms (26 of the defaulted firms have their Z-scores greater than 0). Thus, the Type I error is equal to 18.18%. Similarly, 26.57% of the firms from the non-default group have been classified into the default group. Taken together, 222 firms out of a total of 286 (77.62%) have been correctly classified.

Table 4
Classification results for the matched sample of 286 firms

Correct classification rate (105 + 117)/286 = 77.62%

	Classification Decision		
True State	**Non-default**	**Default**	**Total**
Non- default	105 (73.43%)	38 (26.57%)	143 (100%)
Default	26 (18.18%)	117 (81.82%)	143 (100%)
Total	131 (45.80%)	155 (54.20%)	286 (100%)

The ROC curve for the matched sample is presented in Figure 6. The ROC area is 0.8675 with the asymptotic standard error of 0.021 which in turn leads to the 95% confidence interval that approximately ranges from 0.8264 to 0.9086.

The classification table and the ROC curve discussed above correspond to the matched sample of 286 firms used in the estimation of the model. The performance based on the estimation sample can be viewed as an in-sample result. We can also compute the classification table and the ROC curve for the whole sample of 38,130 firm-year observations. The classification results based on the cutoff value of 0 on the whole sample are given in Table 5. The table reveals that the matched-sample performance of Altman's Z-score model is similar to the performance using the whole sample.

The ROC area for the whole sample is 0.8760 with the asymptotic standard error of 0.0134 which gives rise to the 95% confidence interval of approximately 0.8496 and 0.9023. The ROC area for the whole sample turns out to be close to the one for the matched sample. A lower standard error, however, suggests that the confidence interval for the whole sample is narrower.

Figure 6
The ROC for the matched sample of 286 firms

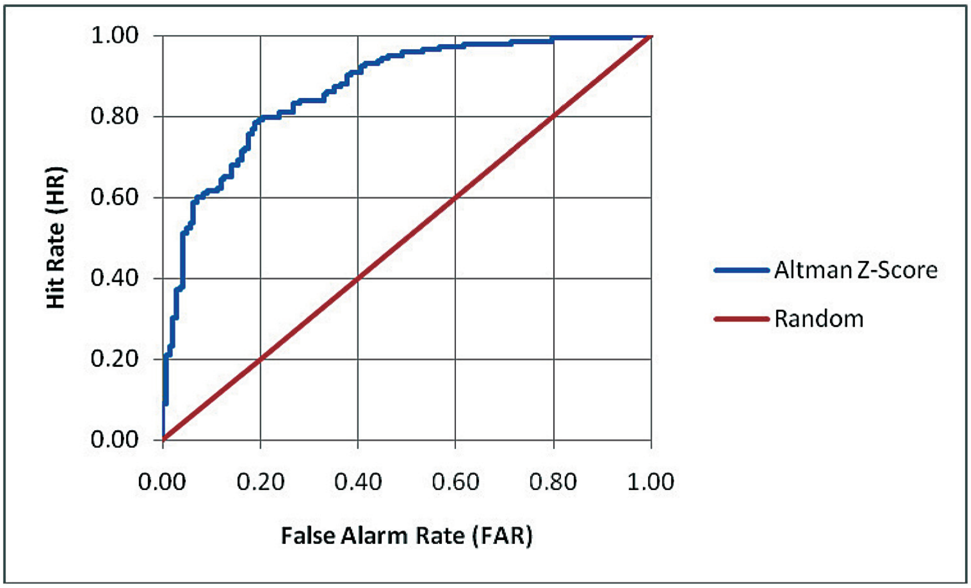


TABLE 5
Classification results for the whole sample (38,120 firm-year observations)

Correct classification rate (28,987 + 117)/38,130 = 76.33%

	Classification Decision		
True State	Non-default	Default	Total
Non- default	28,987 (76.31%)	9,000 (23.69%)	37,987 (100%)
Default	26 (18.18%)	117 (81.82%)	143 (100%)
Total	29,013 (76.09%)	9,117 (23.91%)	38,130 (100%)

II. OTHER CREDIT RATING MODELS

2.1 Ohlson's O-score

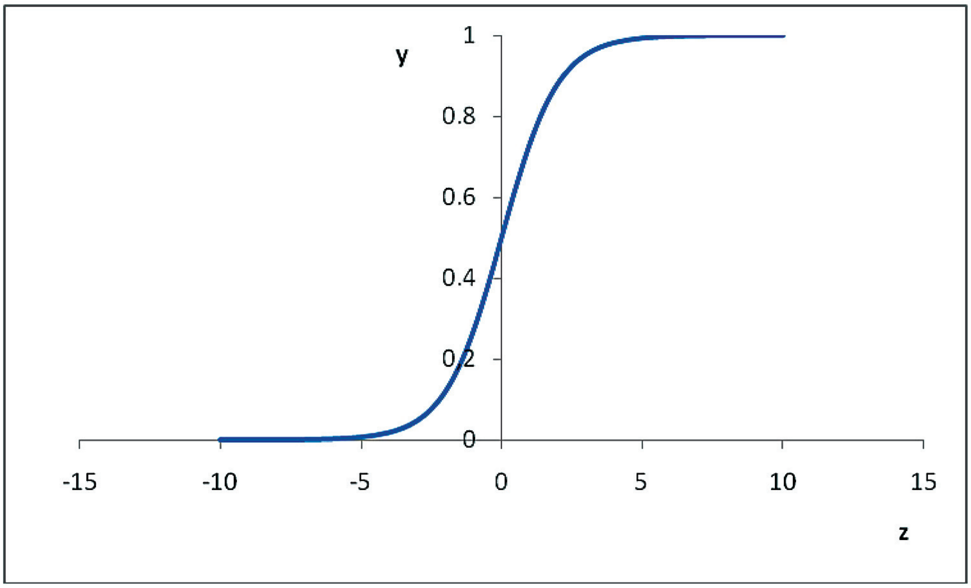
The discriminant analysis for defaults proposed by Altman is useful but does not have the intuitive interpretation of default/survival probability. Moreover, the coefficients in the discriminant function are harder to interpret. In this section, we will introduce logistic regression as a credit rating method that was put forward in Ohlson (1980). Let $P(X_i, \beta)$ represent the probability of a firm defaulting in the next year where $X_i^T = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ are the firm-specific characteristics and β stands for the corresponding set of parameters that define the probability function. Since the parameter values are unknown, they need to be estimated. The firm-specific characteristics can, for example, be represented by the firm's accounting ratios. The odds ratio of default over survival becomes $\frac{P(X_i,\beta)}{1-P(X_i,\beta)}$. In logistic regression, the logarithm of the odds ratio is assumed to be linearly related to the firm's characteristics

$$\log\left(\frac{P(X_i,\beta)}{1-P(X_i,\beta)}\right) = \beta^T X_i = \beta_0 + \beta_i x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \qquad (10)$$

Thus, the probability of default is given as the logistic function of the composite attribute for the i-th firm, i.e., $\beta^T X_i$.

$$P(X_i,\beta) = \frac{1}{1+e^{-\beta^T X_i}} \qquad (11)$$

FIGURE 7
Logistic function $y = f(z) = \frac{1}{1+e^{-z}}$.



The graphic feature of the logistic function is shown in Figure 7. One of the attractive features of the logistic function is the fact that it is bounded between 0 and 1, making it suitable to represent probabilities. The logistic function shown in the figure can be stretched or compressed along the horizontal axis by a linear transformation of z, e.g., by using $y = f(z) = \frac{1}{1+e^{-(a+bz)}}$. In logistic regression, stretching and compression are done with parameter β. The parameters have natural interpretations. For example, a positive coefficient implies that an increase in the value of the corresponding variable will increase the probability of default. Therefore, if we expect the probability of default to decrease with the increase in value of all the variables, we expect all the coefficients to be negative.

The unknown parameters can be estimated using the usual maximum likelihood method which involves choosing the parameter value so that the following log-likelihood is maximized:

$$\underset{\beta}{Max}:\ L(\beta)=\sum_{i\in D}\log\left(P\left(X_i,\beta\right)\right)+\sum_{i\in S}\log\left(1-P\left(X_i,\beta\right)\right) \quad (12)$$

where D is the index set of defaulted firms and S is the index set of survivors. The above log-likelihood function is constructed with the assumption that defaults are independent across firms.

For illustrative purposes, we estimate the logistic regression model using the same five accounting variables as in Altman's Z-score model. The results are given in Table 6.

TABLE 6
Estimation results for the logistic regression on the matched sample of 286 firms

	Coefficient	t-statistic	p-value
X_1	–0.502	–0.70	0.484
X_2	–6.149	–4.85	0.000
X_3	–3.711	–1.17	0.241
X_4	–0.944	–3.24	0.001
X_5	–1.086	–3.27	0.001
Intercept	1.955	5.06	0.000

Unlike the linear discriminant analysis used by Altman's Z-score model, the logistic regression results have easily interpretable parameters. For example, the negative coefficient for the second attribute, i.e., –6.149, implies that this attribute is inversely related to the probability of default. The magnitude of a coefficient is determined by the natural scale of the variable in question. Thus, a direct comparison of the coefficients for different attributes may be meaningless. As all the coefficients corresponding to the attributes considered are negative, it implies that the higher the values of these variables, the lower are their probabilities of default. This result is consistent with the result based on the linear discriminant analysis where a higher value for these variables will make the firm more likely to be classified into the survivors group. Furthermore, logistic regression lends itself to a direct statistical test of significance. For example, even though all the coefficients are negative, the coefficients of X_1 and X_3 are not statistically significantly different from zero.

In the Logistic regression model, we can compute the probability of default for each firm given its characteristics X_i's. As opposed to Altman's Z-scores, default probabilities have direct meaning and suggest how likely obligors are expected to default in the defined horizon. Default probabilities in this context, known as Ohlson's O-scores, can of course be used to rank-order firms. The natural rank orders based on default probabilities run opposite to those of Altman's Z-scores because a safer firm should have a lower Ohlson's O-score but a higher Altman's Z-score. One can generate the CAP and ROC just as in the previous case.

The CAPs for Altman's Z-score model and the logistic regression using the matched sample are presented in Figure 8. Similarly, the ROCs for both models are presented in Figure 9. The accuracy ratio of the logistic regression equals approximately 0.7956 which is higher than 0.7350 for Altman's Z-score model. The ROC comparison is presented in Figure 9. The ROC area is 0.8978 for the matched sample with the asymptotic standard error of 0.019. The 95% confidence interval for the ROC area is thus approximately [0.8605, 0.9350]. The ROC area becomes 0.9028 for the whole sample with the asymptotic standard error of 0.0102,

Figure 8
The CAP for the matched sample of 286 Firms

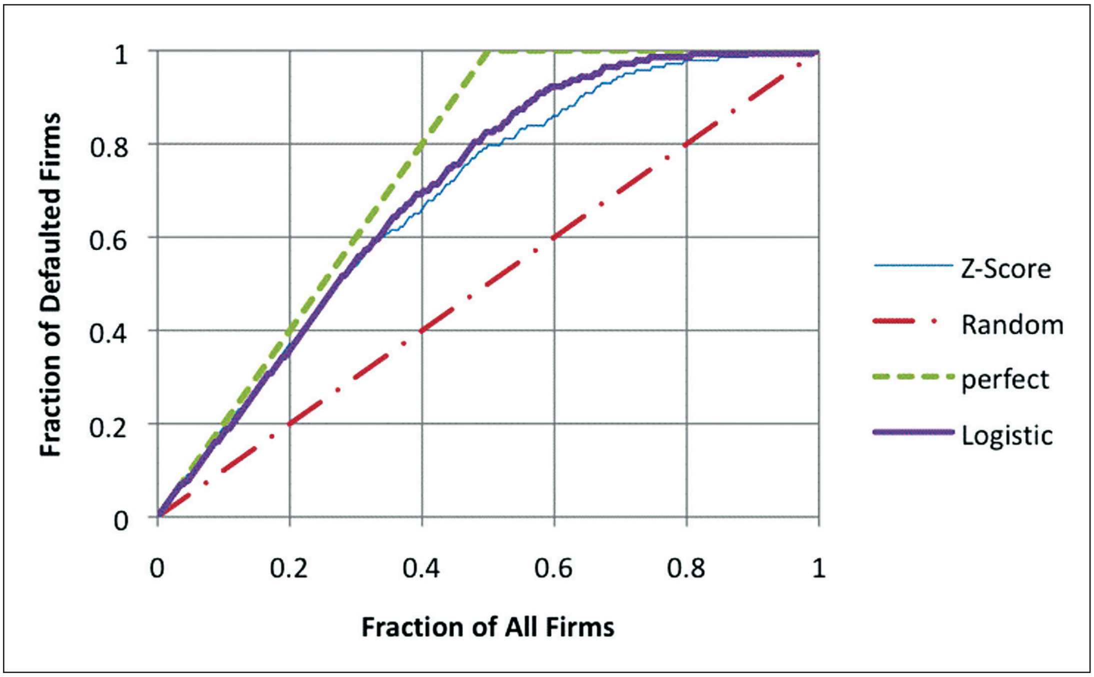
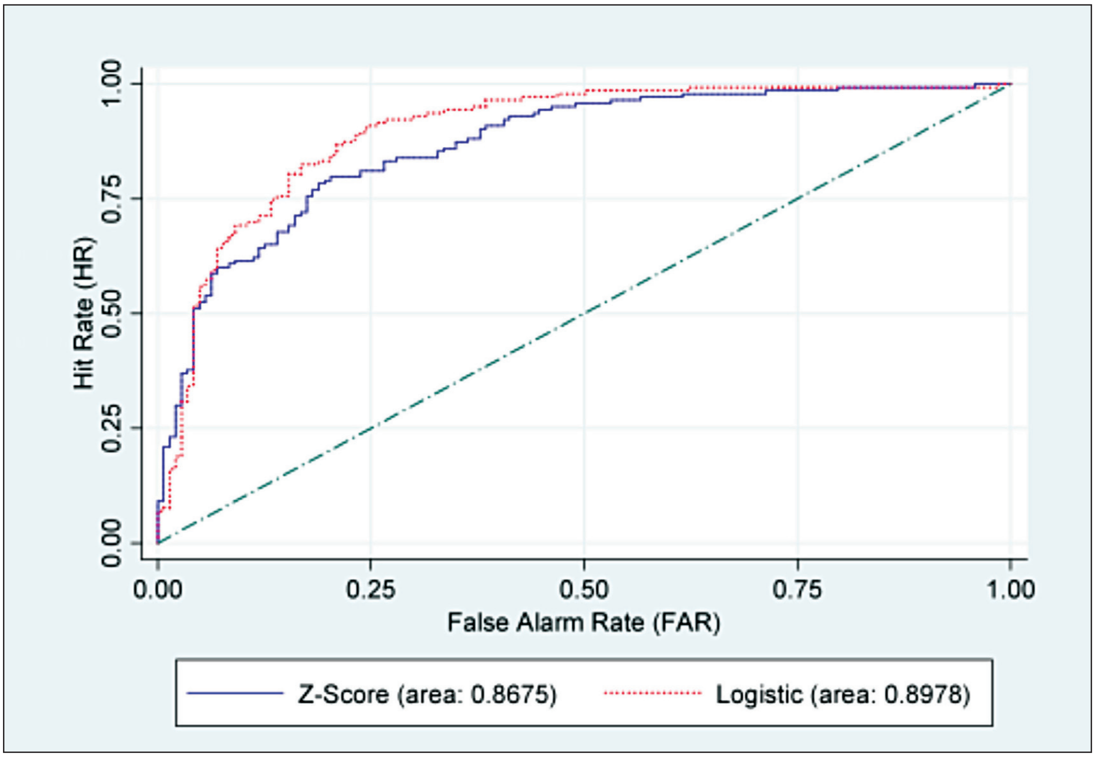


Figure 9
The ROC of the logistic regression for the matched sample



which in turn gives rise to the 95% confidence interval approximately [0.8828, 0.9227]. Even though the ROC areas for the matched sample and the whole sample for the logistic regression are higher than those for Altman's Z-score model, their 95% confidence intervals overlap. Therefore, the improvement may not be statistically significant.

Up until now, the logistic regression model has been implemented with the same set of variables as in Altman (1968). The purpose was, of course, to conduct a controlled comparison of two methods. We now turn to Ohlson's (1980) implementation of logistic regression with the following set of variables:

SIZE = Log(total assets)
TLTA = Total liabilities/Total assets
WCTA = Working capital/Total assets
CLCA = Current liabilities/Current assets
OENEG = 1 if total liabilities exceed total assets, 0 otherwise
NITA = Net income/Total assets
FUTL = Funds provided by operations/Total liabilities
INTWO = 1 if net income was negative for the last two years, 0 otherwise
CHIN = $(NI_t - NI_{t-1})/(|NI_t|+|NI_{t-1}|)$, where NI_t is net income in the most recent period.

As discussed in Ohlson (1980), coefficients of TLTA, CLCA and INTWO are expected to have a positive sign, but the coefficients of SIZE, WCTA, NITA, FUTL and CHIN are expected to be negative. In the case of OENEG, there is no a priori sign for its coefficient. The estimation results are presented in Table 7.

TABLE 7
Estimation results for the logistic regression using Ohlson's variables on the matched sample of 286 firms

	Coefficient	t-statistic	p-value
SIZE (–)	0.024	0.18	0.859
TLTA (+)	8.605	5.91	0
WCTA (–)	1.391	1.12	0.265
CLCA (+)	0.404	1.19	0.236
NITA (–)	–1.323	–0.31	0.759
FUTL (–)	–3.670	–1.23	0.219
INTWO (+)	0.357	0.72	0.469
OENEG (+/–)	–0.575	–0.47	0.638
CHIN (–)	–0.036	–0.13	0.899
Intercept	–7.395	–4.37	0

For the matched sample, some parameter estimates are in general agreement with Ohlson's prediction on the sign but others are not. Among the ten parameters, only two are significant, suggesting that some variables may be redundant and can be removed from the model. The ROC area for Ohlson's O-score model is 0.8894 which is slightly lower than that under the logistic regression implementation using the five variables as in Altman Z-score model. In fact, four parameters are significant with Altman's variables as compared to two under Ohlson's specification. For the matched sample, the asymptotic standard error for the ROC area is 0.0208 which translates to the 95% confidence interval of approximately [0.8486, 0.9301]. When the whole sample is considered, the ROC area is 0.9111 and its 95% confidence interval becomes approximately [0.8900, 0.9322]. The ROC area of 0.9111 is just a touch higher than its corresponding value of 0.9028 when Altman's five variables are used.

2.2 Multi-period logistic regression model

The adoption of logistic regression as in Ohlson (1980) was a methodological improvement. A better or worse accuracy aside, logistic regression offers an intuitive and probabilistic interpretation of default. According to Shumway (2001), however, Ohlson's implementation of logistic regression has inadvertently introduced serious biases into its default predictions. Using one data point per firm essentially ignores the fact that the firm in question has survived some periods prior to that particular time point under consideration. In short, one should use the whole sample so as not to introduce a survival-related bias.

Suppose that one is at time period t. Under logistic regression, the probability that firm i defaults in the next period is $P(X_{it}, \beta)$ whose form was given earlier in Section 2.1. The probability of not defaulting naturally becomes $1 - P(X_{it}, \beta)$. Assume that a firm survives the whole time period from t_0 to T. Then, the likelihood for this surviving firm should factor in the duration of its existence, and becomes $L_i = \prod_{s=t_0}^{T} \left[1 - P\left(X_{is}, \beta\right)\right]$. Similarly, the likelihood for a firm that defaulted in period $t + 1$ should factor in its survival up to time t and the fact that it fails at time $t + 1$. The likelihood thus becomes $L_j = \prod_{s=t_0}^{t-1} \left[1 - P\left(X_{js}, \beta\right)\right] P\left(X_{jt}, \beta\right)$. Then, the likelihood function for the whole sample is given by $L = \prod_{i=1}^{N} L_i$, where N is the total number of firms in the sample. Maximizing L is viewed as a multi-period logistic regression, which explicitly takes into account the fact that data observations on a firm over time actually form a special block. Only by explicitly factoring in such data blocks, can one hope to yield the correct logistic regression estimates.

The estimation results from the multi-period logistic regression model using the whole sample and the set of variables as in Ohlson (1980) are summarized in Table 8. First note that many parameters become statistically significant once we implement the logistic regression correctly. Not all signs are consistent with Ohlson's (1980) predictions, however. For example, the coefficient on WCTA is positive as opposed to the predicted negative value.

The ROC area of the multi-period logistic regression model is 0.7862 with standard error of 0.024, which in turn yields the 95% ROC confidence bounds of 0.739 and 0.8334.

TABLE 8

Estimation results for the multi-period logistic regression model

	Coefficient	t-statistic	p-value
SIZE (–)	0.078	1.46	0.145
TLTA (+)	1.548	3.9	0.000
WCTA (–)	1.420	3.46	0.001
CLCA (+)	–0.020	–1.05	0.295
NITA (–)	–0.239	–1.31	0.190
FUTL (–)	–0.007	–0.16	0.871
INTWO (+)	1.458	7.42	0.000
OENEG (+/–)	1.747	5.09	0.000
CHIN (–)	–0.673	–4.16	0.000
Intercept	–8.097	–12.05	0.000

The shortcoming of Ohlson's (1980) approach lies in the way it is implemented, not in logistic regression itself. The multi-period logistic regression implementation described above has a different implementation shortcoming. A firm can exit from the sample due to a merger/acquisition. Naturally, the probability of surviving a period is not equal to one minus the default probability as in our description. Hence, the multi-period logistic regression as implemented above has also introduced a bias into the system. We will take up the adjustment needed for other forms of exit when we later discuss the Poisson intensity approach to default prediction.

2.3 Artificial Neural Network

The relationship between default and a firm's attributes or common risk factors may be too complex or nonlinear for the modeling approaches such as MDA and logistic regression to handle effectively. A powerful nonlinear modeling tool may capture the default relationship. Artificial neural networks (ANNs) are such a modeling tool. An artificial neural network (ANN) is a mathematical technique used to mimic the way that human brain supposedly processes information. An ANN structure can range from simple to highly complex and computationally intensive. The ANN technique has been widely applied, with varying degrees of success. Its popularity has been aided by the rapid improvement in computing power at a lower cost.

ANN was originally designed for pattern recognition and classification. However, it can also be used for prediction applications. Therefore, it is not surprising to see ANN applied to forecasting bankruptcy; for example, Odom and Sharda (1990), Wilson and Sharda (1994), and Lacher et al. (1995).

An ANN typically comprises several layers of computing elements known as nodes (or neurons).[5] Each node receives input signals from external inputs or other nodes, and processes the input signals through a transfer function, resulting in a transformed signal as output from the node. The transfer function essentially determines how excited a particular neuron is. The transfer function is typically chosen so that a fully excited neuron will register 1 and a partially excited neuron will have some value between 0 and 1. The output signal from the node is then used as the input to other nodes or final result. ANNs are characterized by their network architecture which consists of a number of layers with each layer consisting of some nodes. Finally, the network architecture displays how nodes are connected to one another.

ANN architecture takes a wide variety of forms. Here we will discuss the simple one that has been used for the purposes of bankruptcy prediction.

A popular form of ANN is the multi-layer perceptron (MLP) where all nodes and layers are arranged in a feed-forward manner resulting in a feed-forward architecture. The input layer constitutes the first or the lowest layer of a MLP. This is the layer for the external information. In other words, the MLP receives the external information or input through this input layer. In the context of default prediction, the external information is characterized by the attributes of firms and/or the common risk factors. The last or the highest layer is called the output layer where the ANN produces its result. For default prediction, the output can be thought of as the default probability because a fully excited neuron has the value of 1. In between these two layers, input and output layers, there may exist one or more layers known as hidden layers.

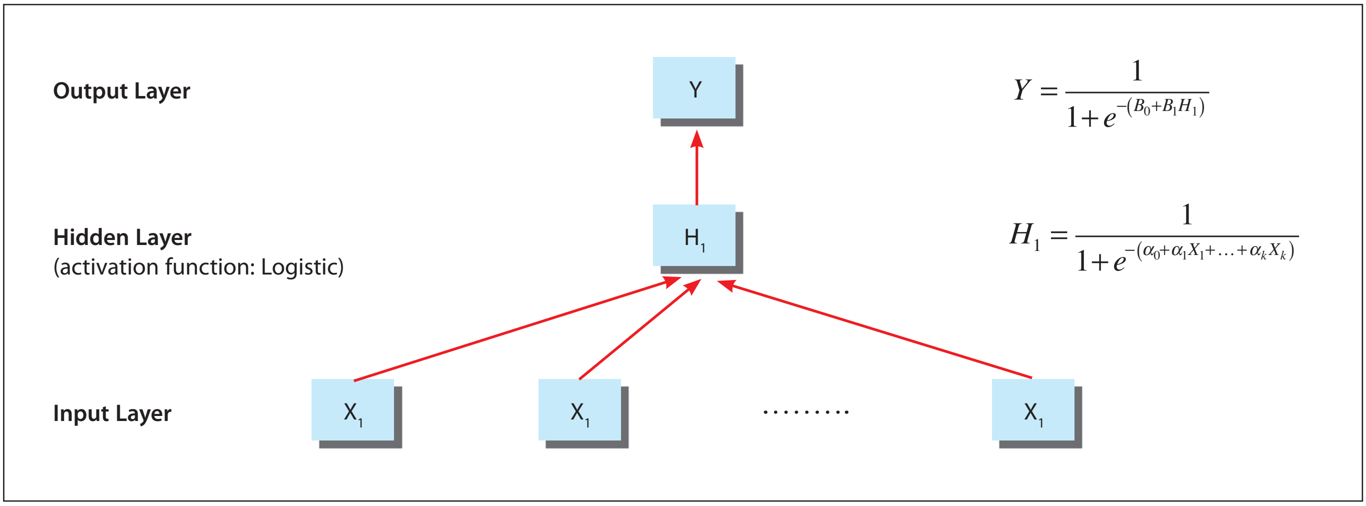
There are almost unlimited variations of the network architecture that represents a MLP. Variations come from the number of hidden layers, from the number of nodes in each layer, and from the ways in which notes are connected. Here we will restrict ourselves to a specific architecture with one hidden layer that will be used later for our default prediction. Since the bankruptcy classification is a two-group classification problem, a three-layer architecture is likely to be sufficient. The three-layer perceptron considered here has only one node in the hidden layer and uses the logistic function, whose value is bounded between 0 and 1, as the transfer function. The exact structure is shown in the figure below.

The input layer consists of the k different inputs, which represents explanatory variables or the firm characteristics. At the hidden layer, the input values, or the activation values of the input nodes, is linearly combined as follows:

$$\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k \qquad (13).$$

In linear regression, the coefficient α_0 is known as intercept. In the neural network terminology, it is known as the *bias* parameter. The linear combination is then translated by the transfer function into an activation

FIGURE 10
A three-layer perceptron architecture


Output Layer
Y
$Y = \frac{1}{1+e^{-(B_0+B_1H_1)}}$
Hidden Layer
(activation function: Logistic)
H_1
$H_1 = \frac{1}{1+e^{-(\alpha_0+\alpha_1X_1+\ldots+\alpha_kX_k)}}$
Input Layer
X_1
X_1
....
X_1

value for the sole node in the hidden layer. As shown in the diagram, the transfer function for the hidden layer is taken as logistic function. Therefore, the activation value of the hidden layer, H_1, is given by

$$H_1 = \frac{1}{1+e^{-(\alpha_0+\alpha_1X_1+\ldots+\alpha_kX_k)}} \qquad (14).$$

The output of the hidden layer is then used as the input to the single node at the output layer. Because we are dealing with a two-group classification situation, a single node is all we need. This differs from the situation of having a single node in the hidden layer where it is a design choice. At the output node, a linear combination of the input, $B_0 + B_1H_1$, is first applied and then followed by applying the logistic function to obtain the output value. Thus, the activation value of the output node, Y, is given by

$$Y = \frac{1}{1+e^{-(B_0+B_1H_1)}} \qquad (15).$$

The activation value or the output of the output layer is the output of the ANN. The output value, Y, is bounded between 0 and 1 because of applying the logistic function. There are two ways of using this output value. One obvious way is to stick to the output value and interpret Y as the likelihood at which an obligor will default. In essence, we have adopted a probabilistic interpretation just like logistic regression. The second possibility is to convert Y to a binary value of 0 or 1. Since we are using the network for the purpose of classification, it may be a sensible thing to do. In the following implementation, we will adopt the second way to convert Y to y as follows:

$$y = \begin{cases} 1 & \text{if } Y \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad (16)$$

and use y to compare with the observed default status (1 for default and 0 for non-default).

The above simple three-layer perceptron is governed by the following two sets of unknown parameters:

$$\alpha_0, \alpha_1, \ldots, \alpha_k \text{ and } B_0, B_1.$$

To suitably determine their values, one needs to "train the network," which means employing a sensible criterion to determine the values for the unknown parameters using a training sample. One intuitive error measure is the mean squared errors (MSE) defined as

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left(a_j - y_j\right)^2 \qquad (17)$$

where a_j represents the target value corresponding to the j^{th} data point and y_i represents the network output using the j^{th} training input values, and N is the number of training sets of input values, i.e. the size of the training sample.

For example, if we use the same set of five ratios to train the network on the matched sample of 286 firms, out of which 143 firms are in the default group. When this set of ratios for firm j is used as input to the above three-layer perceptron, the output of the network, y_j, for this input would be either 0 (representing non-default) or 1 (representing default). The actual default status of a firm is represented by a_j. Therefore, the sample of 286 firms will constitute a training sample that can be used to train the network by finding the value of the parameters that minimizes the MSE.

It should be clear that the training of the network becomes an unconstrained nonlinear minimization problem. One of the most popular algorithms used in applications is backpropogation. This method is a variation of the gradient-based steepest descent method. Naturally, there are other methods of training the network (Zhang et al., 1999). The estimation results are presented in Table 9.

TABLE 9
Results of applying the three-layer perceptron on the matched sample

N		Parameter	Gradient Obj. function
1	α_1	0.144	–4.5E-05
2	α_2	–4.782	–9.5E-06
3	α_3	–0.038	–6.5E-05
4	α_4	–4.145	1.67E-05
5	α_5	–0.907	–3.4E-05
6	α_0	–0.417	–7.2E-06
7	B_1	–5.425	–5.2E-06
8	B_0	2.916	–4.8E-06

The ROC for the matched sample is 0.9067 with the standard error of 0.0188. The result implies that the lower and upper 95% confidence bounds are 0.8699 and 0.9435. Similarly, the ROC for the whole sample is 0.9141 with the standard error of 0.0102 which translates to the lower and upper 95% confidence bounds of 0.8942 and 0.9340.

2.4 Support Vector Machine (SVM)

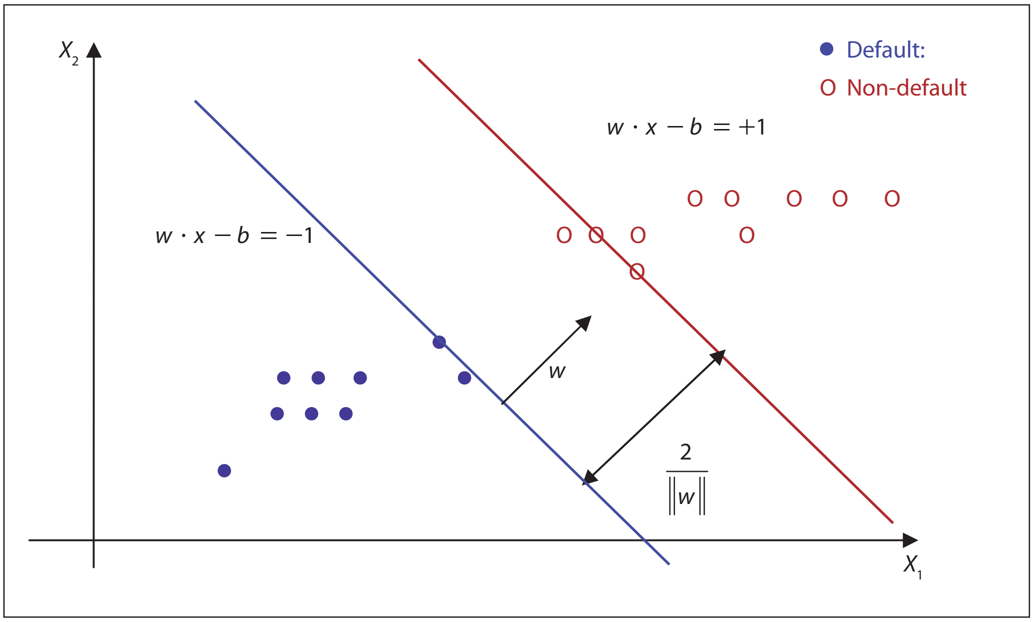
A support vector machine (SVM) is an alternative classification scheme that is quite flexible in dealing with potential non-linear effects within firm's attributes and/or common risk factors. SVM is based on a machine learning technique developed by Cortes and Vapnik (1995). ANN approaches often result in over-fitting data. In other words, ANN can perform well with the training sample due to its flexibility, but perform poorly for data out of the training set. The difference between SVM and ANN is the fact that SVM uses structural risk minimization principle which maximizes the distance to the nearest data points of difference classes in the training data set so as to create a margin for errors when it is generalized to data points outside of the training set. In contract, ANN employs the empirical risk minimization principle that sets out to minimize the error on the training data without creating a margin for errors.

A simple version of the SVM method is a linearly separable case as depicted in Figure 11. Suppose that we have a training sample with firm attributes denoted by a vector x and the firm belongs to one of the two groups denoted by y where $y = -1$ for the default group and $y = +1$ for the non-default group. Then, the classification is conducted with a separating hyperplane defined by $w \cdot x - b$ such that

$w \cdot x_i - b \geq +1$ if firm i belongs to the non-default group
$w \cdot x_i - b \leq -1$ if firm i belongs to the default group

The above conditions can be restated in a more compact form: $y_i \times [w \cdot x_i - b] \geq 1$. The separating hyperplane through +1 and –1 create two parallel hyperplanes as in the diagram. The distance between the two hyperplanes can be shown to equal $2/\|w\|$, where w is a vector perpendicular to both hyperplanes. The objective for

FIGURE 11
Support vector machine (a linearly separable case)



finding the classifier is to find a vector w that maximizes the distance between the two hyperplanes. The solution can be shown to be $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ and the optimal separating hyperplane is given by $\sum_{i=1}^{N} \alpha_i y_i \left(x_i \bullet x\right) - b$ where α_i's and b are the parameters that determine the hyperplane; and, as before, N is the size of the training sample.

When the data is not fully linearly separable, a positive slack variable ξ, is introduced to allow for the classification error for each firm. The set of constraints for different firms can be written as

$$y_i \times \left[w \cdot x_i - b\right] \geq 1 - \xi_i \qquad (18).$$

If one does not want to introduce any slack variable, the original variables x_i's can be nonlinearly transformed to a higher dimensional space, known as feature space, in a way that classification in the feature space become separable. Denote the nonlinear transformation by $\varphi(x_i)$. Then, the optimal separating hyperplane in the feature space would be $\sum_{i=1}^{N} \gamma_i y_i \left(\varphi\left(x_i\right) \bullet \varphi\left(x\right)\right) - b$. The inner product defining the separating hyperplane can be implemented using a kernel function to give rise to:

$$\varpi \cdot \varphi\left(x\right) - b = \sum_{i=1}^{N}{}_{i} \gamma_i y_i k\left(x_i, x\right) - b \qquad (19).$$

Note that γ_i and b are the parameters obtained by maximizing the distance between the two hyperplanes in the feature space. Those vectors corresponding to the non-zero γ_i's are the *support vectors*. A significant number of the coefficients are expected to be zero, however, because their corresponding x_i's are away from the hyperplances. The kernel function $k(x_i, x)$ defines the weight of a data point in the training set x_i to the point of interest x. A kernel function typically assigns declining weights to further away points, and it is not unique. The Gaussian radial basis function is often used as the kernel function in practice; that is,

$$k\left(x_i, x\right) = e^{-\sigma \left\|x_i - x\right\|^2} \qquad (20).$$

The ROC area of the SVM method, for the matched sample, is 0.9406 with standard error of 0.0142. Therefore, the 95% confidence interval is approximately [0.9128, 0.9685]. As for the whole sample, the ROC area is 0.9294 with standard error of 0.0086, which in turn yields the 95% confidence bounds of 0.9126 and 0.9462.

2.5 Poisson Intensity Model

The Poisson intensity model is based on a doubly stochastic process commonly known as the Cox process. Basically, it is a Poisson process whose intensity is a function of other stochastic variables known as covariates. A critical assumption is that the outcome of the Poisson process does not feed back to the covariates that determine the Poisson intensity. In short, two types of stochastic variables have a unidirectional relationship without feedback. A Poisson process can take on many jumps over any finite time period, but the probability for it to make two jumps in one instance is zero. Therefore, one can always kill the Poisson process once it makes the first jump, and the jump time becomes the default time. This idea is often used in the survival analysis. For default modeling, we will present the discussion along the line of the model by Duffie et al. (2007).

Since default is not the only thing that kills a firm, one must also factor in other forms of exit such as merger/acquisition. Failing to factor in the censoring effect arising from other exits will obviously bias default predictions. At the minimum, one thus needs to use two Poisson processes to describe an obligor. Duffie et al. (2007) assumed that the two Poisson processes are independent, conditional on the covariates. In addition, they assumed that the Poisson processes for different obligors are also independent once they are conditioned on the covariates.

Let the intensity be the following function of the covariates and parameters:

$$\lambda\left(x_t; \mu\right) = e^{\mu_0 + \mu_1 X_{1t} + \ldots + \mu_k X_{kt}} \qquad (21)$$

where $\mu_0, \mu_1, \ldots, \mu_k$ are the parameters and ($X_{1t}, X_{2t}, \ldots, X_{kt}$) are the covariates at time t, including firm-specific and macroeconomic variables. Due to the property of Poisson processes, the probability of surviving a small time interval Δt is $e^{-\lambda(x_t;\mu)\Delta t}$. Thus, defaulting in the same interval has the following probability:

$$1-e^{-\lambda(x_t;\mu)\Delta t} \cong \lambda(x_t,\mu)\Delta t \qquad (22).$$

The survival probability over a longer time period can be viewed as surviving many little time intervals, and the end result is the product of all those survival probabilities. Thus, the survival from time 0 to t (divided into n intervals of length Δt) will have the probability equal to

$$e^{-\int_0^t \lambda(x_s;\mu)\,ds} \cong e^{-\sum_{i=1}^n \lambda(x_{(i-1)\Delta t};\mu)\Delta t} \qquad (23).$$

We can define the intensity for other forms of exit in a similar way and denote it by $\delta(x_t;v)$. For a firm to default between t and $t+\Delta t$, it must survive two types of exits over the period of 0 to t and default immediately afterwards. Due to the properties of Poisson processes, the survival portion will be governed by the sum of two intensities. Thus, the probability of default right after time t must be

$$e^{-\sum_{i=1}^n \left[\lambda(x_{(i-1)\Delta t};\mu)+\delta(x_{(i-1)\Delta t};v)\right]\Delta t}\lambda(x_t;\mu)\Delta t \qquad (24).$$

Similarly, the probability a non-default exit between t and $t+\Delta t$ is

$$e^{-\sum_{i=1}^n \left[\lambda(x_{(i-1)\Delta t};\mu)+\delta(x_{(i-1)\Delta t};v)\right]\Delta t}\delta(x_t;v)\Delta t \qquad (25).$$

With the knowledge of the above probabilities, one can construct the likelihood function and carry out the maximum likelihood estimation to obtain the unknown parameters.

Following Duffie et al. (2007), we estimate the Poisson intensity model using the Japanese data set described in earlier sections and focusing on the following covariates:

X_1 = 3-month government bond rate
X_2 = trailing 1-year return on TOPIX
X_3 = distance-to-default
X_4 = firm's trailing 1-year return

The data frequency is monthly (i.e., $\Delta t = 1/12$ years). We have 433,261 firm-month observations in total. The estimation results are presented in Table 10.

Table 10
Result of the Poisson Intensity Model

	Parameter	t-Value
Intercept	−8.109	−24.46
X_1	−1.688	−4.34
X_2	0.647	1.68
X_3	−0.853	−10.84
X_4	−7.200	−13.96

The ROC area for the Poisson intensity model is 0.9738 with the standard error of 0.0077. Therefore, the 95% confidence interval is approximately equal to [0.9586, 0.9890].

III. SUMMARY AND CONCLUDING REMARKS

A range of statistical tools for corporate default analysis has been presented in this paper with the aim of introducing and demonstrating their usage in real default data. We also introduce the cumulative accuracy profile (CAP) and the receiver operating characteristic (ROC) as two intuitive ways of assessing the performance of a credit rating model. From the classical approach of Altman's Z-score to the modern approach of using the Poisson intensity, statistical tools are clearly useful in separating good from poor credit firms.

It should be evident from this review that the modern approach based on the Poisson intensity structure is a far superior method because it is naturally amenable to the panel structure of default data and is also able to easily factor in the data censoring effect arising from other forms of exit such as mergers/acquisitions frequently occurred in realty. In our demonstration, one year was the assumed horizon, but the horizon of interest in default analysis may need to vary for different users.

The Poisson intensity model can in principle, due to its dynamic structure, be aggregated over time to yield a prediction result for any horizon without needing to re-estimate the model. Its conditional independence structure is also amenable to cross-sectional aggregation to form a prediction for a portfolio of obligors. In the parlance of credit risk literature, the Poisson intensity model forms a bottom-up approach.

The Poisson intensity model introduced in this article still has serious shortcomings despite the major advancement offered by its dynamic features. First, it is known to be unable to properly capture the clustered default phenomenon such as is documented in Das et al. (2007). Another limitation is that the time aggregation to different horizons is easy in principle but difficult in reality. The Poisson intensity is a known function of common risk factors and individual firm attributes. For time aggregation to get to a longer horizon of interest, one must prescribe the dynamic processes for all these variables whose future values are unknown. The dimension of the dynamic processes can easily run up to thousands. As an example, we might consider 800 firms, two common risk factors and two attributes for each firm. The total dimension of dynamic processes becomes 1602 (2x800+2). On the aggregation front, a practical solution has been proposed by Duan et al. (2011) and implemented in the RMI non-profit credit rating system which approaches the Poisson model by modeling forward instead of spot intensity.

Notes

[1] See http://www.dnb.com, website of Dun and Bradstreet for more information.

[2] The solution to the problem is given by the eigen-vector corresponding to the largest eigen-value of $W^{-1}B$ where W denotes the within-group cross-product matrix and B the between-group cross-product matrix of the discriminating variables. The (i, j)th element of W and B are respectively given by

$$W_{ij} = \sum_{k=1}^{2}\sum_{n=1}^{N_k}\left(X_{ikn} - \bar{X}_{ik}\right)\left(X_{jkn} - \bar{X}_{jk}\right) \text{ and}$$

$$B_{ij} = \sum_{k=1}^{2} N_k \left(\bar{X}_{ik} - \bar{\bar{X}}_i\right)\left(\bar{X}_{jk} - \bar{\bar{X}}_j\right)$$

where

$\bar{X}_{ik}$ = mean of the ith discriminating variable for group k

$\bar{\bar{X}}_i$ = mean of the ith discriminating variable for the whole sample.

[3] Note that we did not set the intercept to zero as in Altman (1968). The rank orders are not affected with or without intercept, but the interpretation on what level of Z is considered safe depends on the intercept value.

[4] Actually, it can be shown that Accuracy Ratio (AR) defined under the CAP is related to the ROC by $AR = 2.0(ROC - 0.5)$.

[5] Our description of ANN is based on Zhang et al. (1999).

References

Altman, E.I., (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," Journal of Finance, 23, pp. 589–609.

Cortes, C and V. Vapnik, (1995), "Support-Vector Network," Machine Learning 20, pp. 273–297.

Das, S.R., D. Duffie, N. Kapadia and L. Saita, (2007), "Common Failings: How Corporate Defaults Are Correlated," Journal of Finance 62, pp. 93–117.

DeLong, R.E., D.M. DeLong and D.L. Clarke-Pearson, (1988), "Comparing the Area under Two or More Correlated Receiver Operating Characteristics Curves: A Nonparametric Approach," Biometrics 44, pp. 837–845.

Duan, J.C., J. Sun and T. Wang, (2011), "Multiperiod Corporate Default Prediction — A Forward Intensity Approach," National University of Singapore working paper.

Duffie, D., L. Saita and K. Wang, (2007), "Multi-Period Corporate Default Prediction with Stochastic Covariates," Journal of Financial Economics 83, pp. 635–665.

Fisher, R.A., (1936), "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, No. 7, pp. 179–188.

Fitz Patrick, P.J., (1932), "A Comparison of Ratios of Successful Industrial Enterprises with Those of Failed Firms," Certified Public Accountant, pp. 598–605, 656–62, and 727–31.

Lacher, R.C., P.K. Coats, S.C. Sharma, and L.F. Fant, (1995), "A Neural Network for Classifying the Financial Health of a Firm," European Journal of Operations Research, pp. 53–65.

Merwin, C.L., (1942), "Financing Small Corporations in Five Manufacturing Industries, 1926–36," National Bureau of Economic Research, pp. 1–170 (Ph.D. Dissertation, Univ. Pennsylvania).

Odom, M. and R. Sharda, (1990), "A Neural Network Model for Bankruptcy Prediction," Proceedings of the IEEE International Conference on Neural Networks, pp. 163–168.

Ohlson, J.A., (1980), "Financial Ratios and the Probabilistic Prediction of Bankruptcy," Journal of Accounting Research, 18, pp. 109–131.

Richard, M.D. and R.P. Lippmann, (1991), "Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities," Neural Computation 4, pp. 461–483.

Shumway, T., (2001), "Forecasting Bankruptcy More Accurately: A Simple Hazard Model," The Journal of Business, 74, pp. 101–124.

Sobehart, J. and S. Keenan, (2001), Measuring Default Accurately," Credit Risk Special Report, Risk, 14, pp. 31–33.

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

Winakor, A., and R.F. Smith, (1935), "Changes in Financial Structure of Unsuccessful Industrial Companies," Bureau of Business Research, Bulletin No. 51, University of Illinois Press.

Zhang, G., M.Y. Hu, B.E. Patuwo and D.C. Indro, (1999), "Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis," European Journal of Operational Research, 116, pp. 16–32.