A Maximum Likelihood Approach for Reject Inference

Gongyue Chen
Algorithmics Inc.
Toronto, Canada, N2L 3G1
gchen@algorithmics.com

Thomas Åstebro
Rotman School of Management
University of Toronto
tastebro@rotman.utoronto.ca

September, 2005

Abstract

We model reject inference – inferring how a rejected applicant would have behaved had it been granted credit – using a maximum likelihood approach. Though we implement this approach using logit regression the approach is more generally applicable. This reject inference technique embeds the missing data mechanism into the model directly and is therefore expected to be more efficient than other current approaches. We test performance of 5 approaches using real default data. Results show our method to be effective and to improve classification power for credit scoring.

Key Words: reject inference, missing data, missing not at random, maximum likelihood, credit scoring

† Åstebro acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada as well as support from the Canadian Imperial Bank of Commerce.

1. Introduction

Credit scoring represents statistically sophisticated and empirically validated prediction models to assess the risk of providing credit to a person or a business. Credit scoring models are currently relied on by all major banks and financial institutions and are now spreading to other types of organizations and industries. The benefits that credit scoring has provided the banking sector in terms of reduced loan processing costs and diminished aggregate default costs cannot be underestimated.

A reliable sample for credit scoring modeling should be drawn from the population which applies for credit. Reject inference is the process to incorporate into the credit scoring model the information about how a rejected applicant case would have behaved had it been granted the credit. One of the fundamental problems in current reject inference research and practice is possibly the questionable assumption that the parameters distribution of rejects can be inferred directly (or linearly) from accepts. A valid reject inference technique should consider the case that the distributions over the accepted and rejected regions are different.

Two solutions have been proposed. The first is Heckman's bivariate two-stage model. Unfortunately, empirical research of this model shows little promise (e.g., Jacobson and Roszbach, 1999; Banasik, Crook and Thomas, 2003). The second solution is to collect supplementary information about the rejected. For example, Ash and Meester (2002) suggest obtaining credit bureau data on accepts and rejects at the end of the observation period, and using the performance with other creditors over the observation period to infer how the rejects would have performed had they been accepted. There are issues regarding legality, quality of data and costs for obtaining such supplementary information. In general, a major shortcoming of both these methods is an *ad hoc* treatment of data while adding more untested assumptions.

We argue that to improve reject inference further it should be considered a specific case of missing data analysis. The current taxonomy of missing data analysis is based on the work of Rubin (1976), Little and Rubin (1987) and Gelman *et al.* (1995). For any data set, one can define indicator variables R that identify what is known and what is missing. We refer to R as the missingness (Schafer and Graham, 2002). The missing data pattern describes which values are observed in the data matrix and which values are missing. The impact of missing data on the randomness of a sample depends on the process responsible for the disappearance of some sample entries (Sebastiani and Ramoni, 2000), which is usually called the missing data mechanisms. The missing data mechanism concerns the relationship between missingness and the values of variables in the data matrix.

In this research we describe how reject inference can be mapped to missing data analysis. We further argue that reject inference can be considered as a special case of dealing with data missing not at random (MNAR). We extend the approach of reject inference under the frame of MNAR. Our method is based on the traditional maximum likelihood approach. Though we use the logit as an example our method can be applied to other models based on the maximum likelihood principle. We use the logit model simply because it has become the main classification model in credit scoring (Thomas, 2004).

Our reject inference technique embeds the uncertainty of the missing data mechanism directly into a maximum likelihood estimation of the prediction model, and is expected to be more efficient than alternatives. We test the performance of the proposed reject inference technique based on survey data of defaults. To measure model performance we use a model built on the complete sample as benchmark. The benchmark is expected to be optimal. Five reject inference models are estimated and compared with the benchmark: the first model is a Bayesian reject inference model (Chen and Åstebro, 2003); the second is a logit model using the censored sample where all missing data are deleted; the third is Heckman's bivariate two-stage probit model; the fourth is the augmentation model; and the last is based on the logit model with the proposed maximum likelihood reject inference technique. Results show that the proposed reject inference technique is the most efficient among all tested reject inference models and does improve the classification power of the credit scoring model.

The remainder of this paper is organized as follows: Section 2 discusses the mapping of the reject inference problem to a missing data mechanism, while the model is discussed in Section 3. Section 4 outlines the implications and application of this reject inference technique, and compares it to some other techniques. Next, we test the efficiency and power of the method. Then in Section 6 we discuss some related topics, finishing with a brief conclusion.

2 Mapping Reject Inference to a Missing Data Mechanism

The fundamental problem in reject inference is the non-randomness of the sample, a consequence of missing data. This non-randomness removes the grounds for the use of most statistical methods (Corps and Li, 1997). Thus, intuitively, one can think of applying some classical methods of missing data analysis to reject inference. A proper framework for missing data analysis is the missing data mechanism.

The missing data mechanisms are to describe the distribution of missingness, which is regarded as a probabilistic phenomenon (Rubin, 1976). The missing data mechanisms are usually identified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). It is helpful to provide formal mathematical definitions for the missing data mechanisms based on Little and Rubin (1987) before we discuss mapping reject inference problems to missing data analysis.

Let $Y = (y_{ij})$ denote an $(n \times k)$ rectangular data set, with ith row $y_i = (y_{i1}, ..., y_{ik})$ where y_{ij} is the value of variable y_j for subject i. Corresponding to Y, define the missing data indicator matrix $R = (r_{ij})$, such that $r_{ij} = 0$ if y_{ij} is missing and $r_{ij} = 1$ if y_{ij} is present. The missing data mechanism is characterized by the conditional distribution of R given Y, say $f(R \mid Y, \psi)$, where ψ denotes unknown parameters. If missingness does not depend on the value of the data Y, missing or observed, that is, if $f(R \mid Y, \psi) = f(R \mid \psi)$ for all Y and ψ, the data are called missing completely at random (MCAR). Let Y_{obs} denote the observed components of Y, and Y_{mis} the missing components. An assumption that is less restrictive than MCAR is that missingness depends only on the components Y_{obs} of Y that are observed, and not on the components that are missing. That is, $f(R \mid Y, \psi) = f(R \mid Y_{obs}, \psi)$ for all Y_{mis} and ψ. The missing data mechanism is then called missing at random (MAR). The mechanism is called missing not at random (MNAR) if the distribution of R depends on the missing values in the data matrix Y.

For the problems of reject inference, suppose we have a vector X of k independent variables $x_1, x_2, \ldots, x_k$, which is completely observed for each observation[1]. By applying a credit scoring model we assign each observation (credit applicant) i a credit score as $s_i = f(X_i)$ where s_i is the credit score for observation i and $f(.)$ is the scoring function. There is a threshold or cutoff point δ such that when $s_i \geqslant \delta$, credit is granted to observation i, otherwise no credit will be granted. Let a be the indicator variable, and denote $a_i = 1$ if observation i is granted credit, 0 otherwise. Therefore, credit risk, for example credit default/non-default, can be observed if $a = 1$, but is missing if $a = 0$ since this type of information can only be observed for those with credit granted. Denote this observed credit risk as y ($y = 1$ if default, 0 otherwise). Also let's define the missing data of the variable y as $y = ?$. The credit risk y is usually used as dependent variable in the analysis of credit scoring, and hence the missing values in this typical situation are on the dependent variable (outcome). For the sake of simplicity, let's assume that credit scores are bounded in the range $[s_{min}, s_{max}]$. Then the data set can be set in a matrix form as Figure 1 shown:

	x_1	x_2	…	x_k	s	y	a
1					s_{max}	.	*1*
2					.	.	*1*
.					.	.	.
.					.	.	.
.					δ	.	*1*
.					Δ-1	?	*0*
.					.	?	*0*
.					.	.	.
n					s_{min}	?	*0*

Figure 1: Sample matrix form in reject inference

In Figure 1, y is the variable of interest with missing values caused by the credit screening rule where without loss of generality we assume that a credit applicant with a score below δ is rejected.

The pattern of missing data is defined as a univariate pattern where missing values occur on an item with credit score smaller than δ, but a set of (k+1) other items, x_1, x_2, …, x_k, and s, is completely observed. The indicator variable a identifies what is known and what is missing. Therefore the variable a is referred to as an auxiliary variable for the

[1] We assume that a credit application is recorded for all applicants.

missingness (Feelders, 2000). According to Schafer and Graham (2002), we may not have to specify a particular distribution for *a*, but we must agree that it has a distribution. In Figure 1, *y* is the variable of interest with missing values. Denote the complete sample as *A* and partition it as (A_o, A_m) where A_o is the sub-sample with *y* observed (i.e., $a = 1$) and A_m is the sub-sample with *y* unobserved (i.e., $a = 0$). Defined by Little and Rubin (1987), if the distribution of missingness does not depend on A_m, i.e., $P(a \mid A, \psi)$ = $P(a \mid A_o, \psi)$ where ψ is the unknown parameters, missing data are MAR. When the distribution does not depend on A_o either, data are MCAR. If the missing data for reject inference are MCAR, modern missing data theory provides a straightforward solution: one will obtain unbiased estimates by using the complete cases.

Given the univariate pattern for missing values in reject inference, listwise deletion (i.e., delete all observations with missing values) is valid for the MAR missing data mechanism. That is, the parameters of the regression of *y* on any subset of $x_1, x_2, \ldots, x_k$ can be estimated from the complete cases and the estimates are both valid and efficient under MAR (e.g. Graham and Donaldson, 1993). However, this result does not extend to other measures such as correlation coefficients between *y* and *X*, and parameters of the marginal distribution of *y*. But using the complete cases we are guaranteed to establish an unbiased credit scoring model under MAR.

Under MNAR data the problem is quite different. One must specify a model for the missingness that is at least approximately correct. In this case most reject inference techniques may not be valid as they improperly ignore specifying the missingness function. In credit scoring, suppose the model has sufficient classification power. Then it is likely that a missing datum occurs for the reason closely related to the outcome being measured. That is, observations with higher probabilities of default are most likely to have higher probabilities of having lower credit scores. This is a case of a non-ignorable missing data mechanism (Schafer and Graham, 2002; Little, 1995), and this is the case a reject inference technique must be able to handle.

One incorrect mode of thinking is that, given a credit scoring model and a cutoff point, we can always define that $P(a=1|X, y) = P(a=1|X)$; therefore the missing data mechanism for reject inference is always MAR. This is incorrect because, according to Little and Rubin (1987), the probability of missingness should be defined as $P(a \mid A, \psi)$

= $P(a \mid A_o, A_m, \psi)$ where ψ is a parameter vector. In reject inference, ψ will be determined by the credit scoring model as well as the cutoff point δ. If we have $P(a \mid A_o, A_m, \psi) = P(a \mid A_o, \psi)$ for all ψ, then the missing data mechanism for reject inference is MAR. It is likely that for some ψ $P(a \mid A_o, A_m, \psi)$ is not equal to $P(a \mid A_o, \psi)$. Then this is exactly the MNAR case[2].

Notice that as we move from MCAR to MAR to MNAR, the observed values of y become an increasingly select and unusual group relative to the population and the problem of sample selection exacerbates (Schafer and Graham, 2002). Fundamentally in the MNAR case, missing data are generally non-ignorable. In the world of reject inference we therefore have two polar situations, none ideal. One is where the original credit score has no classification power for granting credit. Data are then MCAR and reject inference is not important to apply. However, there is a likely to be a credit quality problem.[3] The second situation is where the original credit score has good classification power causing significant differences in the distributions between the accepted and rejected regions. There is no credit quality problem but a large problem when using the selected sample to update the model.[4]

Assuming that the credit scoring model is well specified and has sufficient classification power we can safely conclude that most reject inference problems can be mapped to a MNAR missing data mechanism. In the following section we describe one such mechanism.

3 Reject Inference in the Logit Model

[2] Let us use an extreme case as an example. Assuming that a credit scoring model perfectly assesses the credit risks at the cutoff point δ, then with this cutoff point the accepted applicants are all good, and the rejected applicants are all bad. If the cutoff point has been decreased so that some applicants with bad risks are also accepted, then the good/bad distributions are clearly different for these two different cutoff points.

[3] Ironically, such problems of ineffective credit scoring models typically appear because the original model was developed on a select sample of accepted applicants.

[4] It is possible that overrides or other selection rules will also be applied to screen applicants. We believe this is just an extended case of MNAR where there are more than one credit score variable. We leave this for future research.

Logit analysis is common in credit scoring, and it is used to linearly estimate the probability for a dichotomous outcome. The logit (as well as probit[5]) models assume that there is an underlying response variable y_i^* defined by the regression relationship $y_i^* = \beta' X_i + u_i$ where X_i is a vector of covariates, β is the vector of coefficients, u_i is the error term and y_i^* is unobservable. We only observe a dummy variable y defined by $y = 1$ if $y_i^* > 0$ and $y = 0$ otherwise (Maddala, 1983). We get $Prob(y_i = 1) = Prob\ (u_i > -\beta' X_i) = 1 - F(-\beta' X_i)$ where F is the cumulative distribution functions for u_i. The related likelihood function is

$$L(\beta) = \prod_{y_i=0} F(-\beta' X_i) \prod_{y_i=1} [1 - F(-\beta' X_i)] \quad (1)$$

and its log likelihood function is

$$l(\beta) = \sum_{y_i=0} \ln[F(-\beta' X_i)] + \sum_{y_i=1} \ln[1 - F(-\beta' X_i)] \quad (2).$$

The logit model assumes the cumulative distribution of u_i to be the logistic. That is,

$$p_i(y=1) = F(-\beta' X_i) = \frac{\exp(-\beta' X_i)}{1 + \exp(-\beta' X_i)} = \frac{1}{1 + \exp(\beta' X_i)} \quad (3)$$

where $p_i(y = 1)$ is the estimated probability of $y = 1$ for observation i.

In this model the dependent variable y, with the value either 0 or 1, is completely observed. However, in credit scoring this variable is unobserved for applicants who have been rejected. To perform reject inference with the logit model, following the logic of the above section we define that reject inference can be mapped to the missing data mechanism of missing not at random. This requires a method for handling non-ignorably

[5] We are not likely to get very different results using the probit and logit model since the cumulative normal distribution and the logistic distribution are very close to each other, except at the tails (Maddala, 1983). Hence, in this paper we focus on the inference of logit model simply because it has a closed form expression for the cumulative distribution.

missing outcomes in binary data, while covariates are completely observed. Define λ_i to be the probability of missingness for the observation *i* where the credit quality is unobserved. Also define $y = 1$ as bad risk and $y = 0$ as good for the observation where the credit quality is observed. Therefore, the sample used for reject inference contains two types of observations. One type is those observations where *y* is observed (e.g., $y = 0$ or $y = 1$), and the second type is those observations where *y* is missing. Consequently, for a sample with missing data in *y*, using the expectation for the log likelihood function equation (2) becomes

$$\begin{aligned} l(\beta) = & \sum_{y_i=0} \ln[F(-\beta' X_i)] + \sum_{y_i=1} \ln[1-F(-\beta' X_i)] + \\ & \sum_{y_i=missing} \{(1-\lambda_i)\ln[F(-\beta' X_i)] + \lambda_i \ln[1-F(-\beta' X_i)]\} \end{aligned} \quad (4).$$

Using equation (4) the reject inference problem simply becomes how to estimate the missing data mechanism. However, the estimation of the missing data mechanism is data dependent. To apply the proposed model for the simple case that the credit screening policy is solely based on a scoring model, we propose not to posit strong theory about the missingness mechanism; rather, we suggest using internal and external information to extrapolate the probability of being missing. An example of the external information is the good/bad distribution obtained from the modeling process of the original scorecard on which the credit screening decision is based. On the other hand, the internal information is anything contained in the design sample. Since we assume that the credit score has sufficient classification power, the probability of being bad in the observed groups can also be used to infer the missingness mechanism. Some possible methods for this inference include regression, linear extrapolation or exponential extrapolation. Finally, a better solution may be to estimate the missingness mechanism by computing a weighted average of group specific estimates using both external and internal information. Using combined external and internal information is superior to using one information source since one can always assign a weight of 1.00 to one information source[6].

[6] In the following application we assume that external and internal information have equal weights.

4 Implications of the Model

If the missing data mechanism is missing at random, the corresponding model for this approach is the EM algorithm (Dempster, Laird and Rubin, 1977.) Since missing data are missing at random for the EM algorithm, the expected probability of missingness can be estimated from the design sample. If the sample size used for modeling is sufficiently large, the missing at random assumption implies that list-wise deletion, where all missing data are dropped, is efficient for reject inference. However, if the missing data mechanism is missing not at random, a model using the method of list-wise deletion will be biased and reject inference should be applied. Furthermore, unlike the EM algorithm, we cannot use the expected probabilities estimated from data because the available sample is not representative for the entire population. Instead, equation (4) gives the correct log likelihood function in the case of missing not at random.

Paik, Sacco and Lin (2000) suggest a computational approach comparable to this model (i.e. equation (4)). Their original model is aimed at dealing with bivariate binary data with nonignorably missing outcome. Their approach can be modified for reject inference with the logit model. First, let y_i be the binary outcome of credit risk indicating whether the ith observation is bad (=*1*) or not (=*0*). Also let $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ be a vector of covariates for observation i, and the number of covariates is k. From equation (3) we have $p_i = F(-\beta' X_i) = \frac{1}{1+\exp(\beta' X_i)}$ where β is the vector of the parameters of interest. Let r_i be the observation indicator of y_i, and assume there are n observations. If all data are completely observed, the estimates of β can be obtained simply by solving (Paik et al., 2000)

$$\Sigma_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} (y_i - p_i) = 0 \qquad (5).$$

The strategy for handling missing y_i is to replace it with its conditional expectation given the observed data, $E(y|X_i, r_i, s_i)$ where s_i is the credit score of the observation i, and solve

$$u = \Sigma_{i=1}^{n}\{r_i \begin{pmatrix} 1 \\ x_i \end{pmatrix} (y_i - p_i) + (1 - r_i) \begin{pmatrix} 1 \\ x_i \end{pmatrix} \times [E(y_i \mid x_i, r_i, s_i) - p_i]\} = 0 \qquad (6).$$

In equation (6) $E(y|x_i, r_i, s_i)$ is the expected probability of missingness λ in equation (4). Suppose $E(y_i|x_i, r_i, s_i)$ is available, Paik *et al*. (2000) then use the Newton-Raphson algorithm to compute estimates for equation (6), and a jackknife variance estimate (e.g., Lipsitz *et al.*, 1994) for the computation of variance.

The model we propose here should not be inferior to the model proposed by Paik *et al*. (2000) since the estimates of the maximum likelihood method are asymptotically most efficient Modern statistical software (e.g. STATA) also eliminates the disadvantage of computational demand for the maximum likelihood approach, and makes the program work much easier for our model. The critical issue for successful application of this model is to estimate the probability of missingness λ in equation (4).

5 Test

Using the 1993 and 1998 National (U.S.) Surveys of Small Business Finances (NSSBF) data sets[7] we design an experiment to test the power and efficiency of the proposed maximum likelihood based reject inference technique. First, we develop a credit scoring model to predict the probability of credit delinquency using the complete 1993 NSSBF sample. Next, applying that credit scoring model we simulate credit granting procedures for the 1993 NSSBF sample (within-sample test) and 1998 NSFFBF sample (out-of-sample test) so that both samples contain (simulated) credit acceptance/reject information. The consistency applied in the survey design of the 1993 and 1998 NSSBF made all variables used in the credit scoring model based on the 1993 NSSBF available for the 1998 NSSBF. Therefore, these two data sets are ideal for this simulation test. Using these two samples with simulated credit acceptance/rejection information we compare the performance of the proposed reject inference technique with other techniques. In the following we summarize the statistical procedures for this simulation test, but will discuss these procedures further in the next sub-sections.

To investigate the robustness of the proposed reject inference model we simulate two credit granting policies by applying two cutoff scores so that the degree of missingness is different across the two selected samples. This also means that where there is potential sample selection bias the degree of bias will be different. A cutoff score is selected such that the acceptance rate is equal to the good rate (or one minus bad rate) of the complete 1993 and 1998 NSSBF samples.[8] For convenience we name this case "weak selection." The second case is "strong selection" where the cutoff score was chosen so that comparing to the "strong selection" case approximately twice the number of applicants was rejected.

Using the observed bad rates as the proxy for the missingness function, we can provide estimation for the missing data mechanism. For within-sample test (test using the 1993 sample), the missingness function is inferred from "internal" information of the bad rate in the selected 1993 sample. This scenario represents the situation that, except for the design data, there are no other sources of information that can be used to infer missingness function. For out-of-sample test (test using the 1998 sample), there are at least two possible sources of information for missingness inference. The first source is certainly the sample used for out-of-sample test. We term this source "internal" information. The second source is the sample used to develop the original credit scorecard. We term it "external" information. In this example, the "external" information for out-of-sample test is the 1993 sample. In reality it is not uncommon to obtain this type of "external" information. For instance, vendors of scorecards generally provide clients the bad rate distributions of their design data. In out-of-sample test the missingness mechanism is computed as a weighted average of both external and internal information[9].

Proper estimation of the missing data mechanism allows us to implement the proposed reject inference model. In this test we assume a strict credit screening policy based on the credit score so that there is no additional selection rule. For a score higher

---

[7] For more details please see http://www.federalreserve.gov/pubs/oss/oss3/nssbftoc.htm. We assume that the samples obtained from the Surveys are representative of the true distributions of the underlying universes.

[8] The good rate is the proportion of good accounts among those accepted for credit.

[9] Here the selection of equal weights for external and internal information is arbitrary. In reality the weighting system is determined by the belief to what degree the missingness can be represented by these two information sources. A cautious procedure may be to first analyze and compare the distributions of external and internal information.

than the cutoff point the credit application is denied and credit quality is unobservable. Unobservable credit qualities are indicated as missing. The expected missingness probabilities are assigned to those unobserved credit qualities. The details of model performance will be described in section 5.6.

In sum, the following is the simulation test steps:

1. Build a credit scoring model using 1993 NSSBF data;
2. Apply the model from step 1 to 1993/1998 NSSBF samples to generate simulated samples with credits rejected. Note that there are two types of simulation, weak and strong selections, and there are a total of four simulated samples for model testing.
3. Estimate the missingness probabilities for all four simulated samples from step 2.
4. Based on the results from step 3, apply the proposed reject inference technique to created simulated credit scoring models, and test the performance of the model by comparing with other reject inference techniques.

In the following we give detailed description for above procedures including the model performance measurement used (Section 5.1), the data samples (Section 5.2), original credit scoring model (Section 5.3), the generation of the simulated samples (Selection 5.4), estimation of missing data mechanisms (Section 5.5) and model performance (Section 5.6).

5.1 Performance Measurement

Measuring performance of a credit scoring model is a key issue. We are interested in how the classification power of a model for the targeted population distribution is affected by missing data and potential corrections for missing data. For this we use four measures. The first is the Kolmogorov-Smirnov (KS) test that is commonly applied in industry to validate the power of a model in separating two groups (e.g., goods and bads) from each other. Another group of performance measures deals with the proportion of cases that are misclassified. To simplify the presentation we use the bad rate which is computed simply as the proportion of observed "bad" among those accepted. To apply bad rate measure we set the classification rule such that the acceptance rate (e.g., the rate to grant an applicant a credit) is equal to the proportion of good in the population. However, the within-sample bad rate is expected to be an optimistically biased estimate of future out-of-sample performance (Hand 1998). Some more appropriate measures for (in)accuracy are Brier score and logarithmic score. These two measures are also the measures for forecasting errors, and used to measure out-of-sample accuracy[10].

5.2 Data

The 1993 and 1998 NSSBF databases are used for this simulation test. The NSSBF surveys collected information on the availability of credit to small and minority-owned business. The surveys provide detailed information on firms' characteristics, owners' characteristics, business performance and the firms' financial status. The 1993 NSSBF database contains 4637 observations where all enterprises operated under the current ownership during 1992 with fewer than 500 full-time equivalent employees[11]. To make the sample applicable for potential bank financing we drop observations where the firm has zero total expenses or zero total liability in 1992. The development sample size is therefore 4589 observations, in which 20% of the population has been delinquent on at least one business obligation within three years. The 1993 NSSBF data will be used to build a credit scoring model and test the proposed model as with-in-sample test.

The 1998 NSSBF database will be used for out-of-sample test in this simulation test. The 1998 NSSBF data contains 3561 observations of small businesses that were in operation during December 1998. The survey structure is similar to that of the 1993 NSSBF. Applying the same sampling procedure as for the 1993 sample, the final size of the test sample is 2805, in which 493 (17.6%) of these observations have been delinquent on at least one business obligation.

5.3 Credit Scoring Model

Based on the 1993 sample we create a credit scoring model using logit regression to predict the probability of delinquency. We define the dependent variable $y = 1$ if a firm has been delinquent on at least one business obligation within three years, and 0 otherwise. Definitions of variables are presented in Appendix 1. Appendix 2 displays sample statistics and Appendix 3 presents estimation result.

---

[10] Hand (1997) provides a good review of classification rules. Hand (2000) suggests that common measures of precision are the Brier score and the logarithmic score.

[11] The database excludes agricultural enterprises, financial institutions, not-for-profit institutions, government entities, and subsidiaries controlled by other corporations.

Using the logit regression result we calculate the predicted probability of being delinquent for each observation. The credit score is defined as this predicted probability times 1000 with a range from 0 to 1000. The smallest observed score is 0 and the largest is 865. The KS for the model sample (the 1993 sample) is 0.35. We simulate an acceptance rate of 80%. The cutoff score δ representing this acceptance rate is 297. Based on this cutoff score we obtain a selected sample where the bad rate is 14.2%. By randomly selecting 80% of the applicants the expected bad rate will be 20%. Therefore, the credit score will improve credit quality by 29% over random choice. The Brier score for the model is 0.28 and the logarithmic score is 0.44.

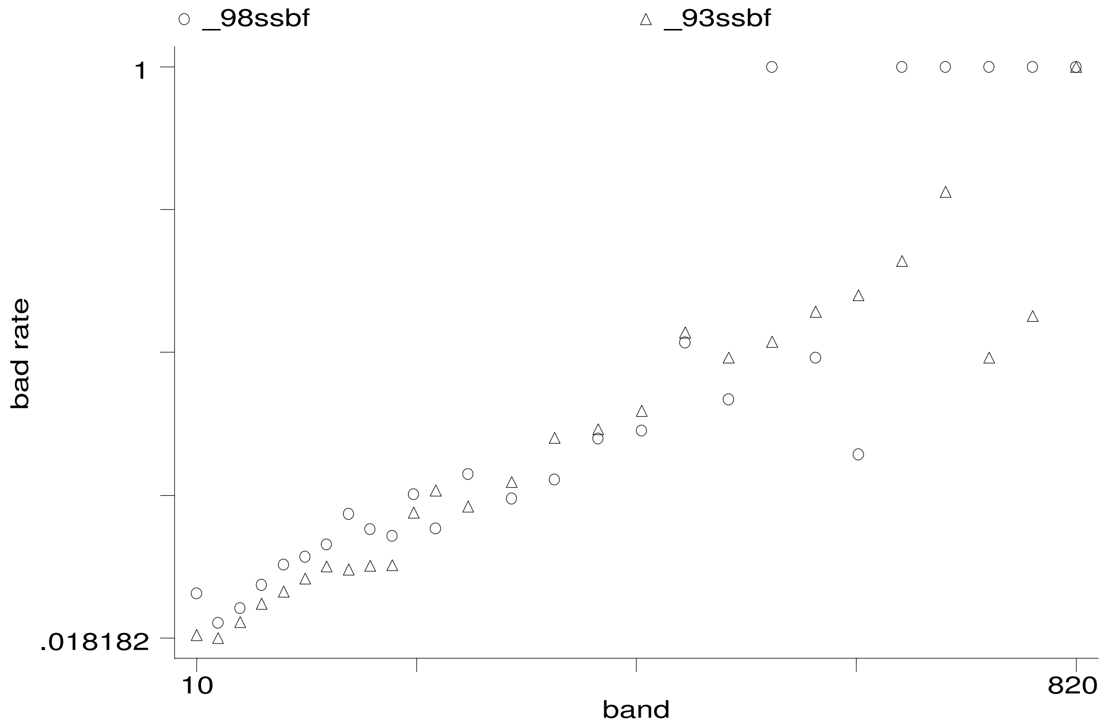
We further test the performance of the original scoring model out-of-sample using the 1998 NSSBF. The KS statistic for this sample is 0.23, which represents a decrease of 36.4% compared to the development sample. If we choose the cutoff so that 493 observations would be rejected, the cutoff score is 218 and the bad rate is 15.1%. Although this bad rate implies that applying the credit score will improve the credit quality by 14% over random selection, it is clear that the classification power is weaker in out-of-sample tests. The Brier and logarithmic scores are 0.28 and 0.46 respectively, somewhat higher than those obtained for the development sample. This further indicates that the populations are different and that re-development of the scoring model may be necessary to maintain classification accuracy. Besides the reason that the worse performance of out-of-sample test is caused by population drift over time, the original model may be biased for the 1998 sample. Many variables that are significant in the 1993 sample are not significant for the 1998 sample. Analysis shows that the variables significant for the 1998 sample model are quite different from those significant in the original credit scoring model based on the 1993 sample.

Before removing outcome data that are screened out due to the credit policy and for illustrative purposes we graph the bad rates for each score band for both the development and selected samples in Figure 2.[12] As seen, the bad rate is increasing pretty

[12] We split the credit score into 31 bands. For the score range from 0 to 240, we use 12 bands, all with 20 as band width. For the score range from 241 to 1000, we use 19 bands all with 40 as band width. The starting range is from 0 to 20 so that the value of the band width is actually 21 for the first band. Also note that the bad rates in some bands are 100%. One reason is that there are few observations in these bands and they are all bad. On the other hand, in some bands there are no observations at all and the bad rates are assumed to be 100%. This appears to be a harmless assumption since no imputation will be made for these bands.

much linearly with the scores for both the development and test samples. However, there is larger variation of bad rates at higher score bands in both samples which is driven by the smaller number of observations within each band at higher score bands.

Figure 2: Bad Rate Distributions for the Development and Test Samples



5.4 Simulation for Samples with Rejected Credit

By applying the credit scoring model described in the above section to 1993 and 1998 NSSBF samples, we can simulate samples with rejected credit for testing of reject inference techniques.

We apply two credit screening policies: weak and strong selection. Setting the acceptance rate equal to the good rate in the complete 1998 sample the cutoff score 218 is obtained. The result shows 17.6% or 493 observations have been screened out. The second case is "strong selection" where the cutoff score is 160 such that 33.3% or 934 observations have been screened out.[13]

[13] Since the sample size is not large enough, and credit scores are not evenly distributed, the amounts of the rejected applicants for "strong selection" case are only approximately double of those for "weak selection" case.

To make sure that the assumption of missing not at random is valid, we also perform within-sample tests. For the case of weak selection we drop 20% (or 918 observations) of the 1993 sample based on the credit score. The cutoff score for this acceptance rate is 297. For the strong selection case, we select the cutoff score 200 such that 38.5% (or 1767 observations) of the 1993 sample is dropped.

5.5 Estimation of the Missing Data Mechanism

Upon obtaining the simulated samples with rejected credit, we need to estimate the missing data mechanism (i.e., the λ in equation (4)). As discussed, the missing data mechanism can be represented by estimates of the bad rate. To simplify the problem we then estimate the bad rates for those credit rejected based on the score bands using the internal and external information sources. First, the development sample provides a complete distribution of bad rates conditional on credit score bands ("external" information). Second, for the sample with credit screening applied, information of bad rates in the accepted region is available ("internal" information). This information can be extended to estimate the bad rate distribution in the rejected region. Linear regression is applied since Figure 2 indicates that there are strong linear trends in bad rate distributions.

Table 1 presents estimations for missing data mechanisms of the 1993 and 1998 samples [cols. (4), (7), (8) and (9)] over the score bands [col. (1)]. There are two missing data selection biases for each sample: weak and strong selection. To estimate the missing data mechanism for the 1998 sample we use two information sources that are equally weighted. External information contains the bad rate distribution obtained from the 1993 sample [*e.g.,* col. (2)][14]. That is a bad rate distribution obtained by applying the credit scoring model to the 1993 sample. Internal information contains the bad rate distribution of the accepted region in the 1998 sample [*e.g.,* col. (3)]. The bad rate distribution for missing data score bands is predicted using a regression model based on data from the accepted region. That is, for example, in case of column (3) we estimate a linear regression model using the 1998 sample with score 0 to 160, then predict the probabilities being bad using the estimation of the regression model for scores 161 to 1000. For the

[14] Since data on bad rates for the 1993 sample for scores over 721 are scarce a regression model is used to predict these values.

final estimation of the missing data mechanism, we weighted the estimations of "external" and "internal" information equally. For the 1993 sample we assume the only available information is the bad rate distribution from the accepted region of the 1993 sample, and based on this available information linear regression models are obtained to predict the missing data mechanism. Details are presented in the Notes of Table 1. Note that using linear regression models to predict the expected probability of being bad is purely an empirical practice. We tested some other approaches such as log-linear regression model, but feel that the linear regression model is better to fit in this case. We have to admit that there may be other suitable approaches that are worthy of trying.

5.6 Performance

Proper estimation of the missing data mechanism allows us to implement the proposed reject inference model (i.e., using equation (4) in Section 3). We assume a strict credit screening policy based on the credit score so that there is no additional selection rule. For a score higher than the cutoff point the credit application is denied and credit quality is unobservable. Unobservable credit qualities are indicated as missing. Based on Table 1, the estimated missingness probabilities are assigned to those unobserved credit qualities.

To measure model performance, we use a model built on the complete sample as benchmark. The benchmark is expected to be optimal. Five models are estimated and compared with the benchmark: the first model is the Bayesian bound and collapse model (Chen and Åstebro, 2003); the second one is the logit model using the censored sample where all missing data are deleted; the third is the bivariate two-stage probit model; the forth is the augmentation model; and the last is based on the logit model with the proposed maximum likelihood reject inference technique. The performance of these five models is expected to be inferior to the benchmark. What we are interested in is their relative performance compared to the benchmark. Model performance measures are computed for all these models as well as the benchmark models for comparison. Table 2 displays results of the 1998 NSSBF sample (out-of-sample test), and Table 3 is the results of the 1993 NSSBF sample (within-sample test).

Table 1: Estimation for Missing Data Mechanisms.

Band	*1998 NSSBF sample*						*1993 NSSBF Sample*	
	Strong Selection			*Weak Selection*			*Strong Selection*	*Weak Selection*
	1993 NSSBF bad rate	*1998 NSSBF bad rate*	*Missing data mechanism*	*1993 NSSBF bad rate*	*1998 NSSBF bad rate*	*Missing data mechanism*	*Missing data mechanism*	*Missing data mechanism*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
161-180	14.19%	23.26%	18.72%					
181-200	14.34%	25.56%	19.95%					
201-220	23.42%	27.87%	25.64%	23.42%	26.52%	24.97%	18.15%	
221-240	27.12%	30.17%	28.64%	27.12%	27.07%	27.09%	19.73%	
241-280	24.41%	33.63%	29.02%	24.41%	29.98%	27.20%	22.10%	
281-320	28.63%	38.23%	33.43%	28.63%	33.87%	31.25%	25.26%	28.63%
321-360	36.22%	42.84%	39.53%	36.22%	37.75%	36.98%	28.42%	33.21%
361-400	37.72%	47.45%	42.59%	37.72%	41.64%	39.68%	31.58%	37.08%
401-440	40.87%	52.06%	46.46%	40.87%	45.52%	43.20%	34.74%	40.94%
441-480	54.32%	56.67%	55.49%	54.32%	49.41%	51.86%	37.90%	44.81%
481-520	50.00%	61.27%	55.64%	50.00%	53.29%	51.65%	41.06%	48.68%
521-560	52.73%	65.88%	59.30%	52.73%	57.18%	54.95%	44.22%	52.54%
561-600	57.89%	70.49%	64.19%	57.89%	61.06%	59.48%	47.38%	56.41%
601-640	60.71%	75.10%	67.91%	60.71%	64.95%	62.83%	50.54%	60.27%
641-680	66.67%	79.71%	73.19%	66.67%	68.83%	67.75%	53.70%	64.14%
681-720	78.57%	84.31%	81.44%	78.57%	72.72%	75.65%	56.86%	68.01%
721-760	76.30%	88.92%	82.61%	76.30%	76.60%	76.45%	60.02%	71.87%
761-800	80.45%	93.53%	86.99%	80.45%	80.49%	80.47%	63.18%	75.74%
801-840	84.61%	98.14%	91.37%	84.61%	84.37%	84.49%	66.34%	79.61%
841-880	88.77%	100.00%	94.38%	88.77%	88.26%	88.51%	69.50%	83.47%
881-920	92.92%	100.00%	96.46%	92.92%	92.15%	92.53%	72.66%	87.34%
921-960	97.08%	100.00%	98.54%	97.08%	96.03%	96.55%	75.82%	91.21%
961-1000	100.00%	100.00%	100.00%	100.00%	99.92%	99.96%	78.98%	95.07%
Weight	0.5	0.5		0.5	0.5			

(1) A regression model is used to estimate the bad rates in the range 721 –1000 for the 1993 sample.
Bad rate = -0.00589 +0.001039*score + error
(-0.557) (36.017)
where value in ( ) is the *t* ratio. The R square is 0.983, and adjusted R square is 0.983.

(2) In the case of weak selection for 1998 NSSBF sample, the regression model to estimate the bad rates of the rejected region of 1998 NSSBF sample is
Bad rate = 0.04728 + 0.0009713*score + error
(2.909) (7.583)
The R square is 0.865, and adjusted R square is 0.850.

(3) In the case of strong selection for 1998 NSSBF sample, the regression model to estimate the bad rates of the rejected region of 1998 NSSBF sample is
Bad rate = 0.03673 + 0.001152*score + error
(1.975) (5.712)
The R square is 0.919, and adjusted R square is 0.845.
In the case of weak selection, the bad rate for band 201-220 is estimated from the observed data directly.

(4) In the case of weak selection for 1993 NSSBF sample, the regression model to estimate the bad rates of the rejected region of 1993 NSSBF sample is
Bad rate = 0.003456 + 0.0009666*score + error
(0.286) (13.292)
The R square is 0.936, and adjusted R square is 0.931.

(5) In the case of strong selection for 1993 NSSBF sample, the regression model to estimate the bad rates of the rejected region of 1993 NSSBF sample is
Bad rate = 0.01558 + 0.00079*score + error
(1.504) (8.792)
The R square is 0.906, and adjusted R square is 0.894.

Table 2: Model Performance for the 1998 NSSBF Sample.

		Weak selection					*Strong selection*				
	Optimal	Modified logit	Bayesian	Bivariate	Censored	Augmentation	Modified logit	Bayesian	Bivariate	Censored	Augmentation
KS	0.361	0.349	0.355	0.347	0.35	0.344	0.333	0.327	0.337	0.353	0.341
Bad rate	12.5%	12.7%	12.5%	12.4%	12.3%	12.5%	13.1%	13.0%	13.0%	12.6%	12.9%
Brier score	0.251	0.253	0.254	0.255	0.252	0.253	0.260	0.261	0.264	0.255	0.259
Difference test 1		2.461	2.498	2.754	1.717	2.237	4.271	4.463	4.726	2.796	3.862
		[0.014]	[0.013]	[0.006]	[0.086]	[0.025]	[0.000]	[0.000]	[0.000]	[0.005]	[0.000]
Difference test 2			0.508	0.947	-0.935	-0.081		1.748	1.688	-2.159	-0.198
			[0.612]	[0.344]	[0.350]	[0.936]		[0.081]	[0.092]	[0.031]	[0.843]
Logarithmic score	0.402	0.407	0.408	0.409	0.405	0.407	0.416	0.418	0.424	0.409	0.416

1. Optimal = model based on complete sample; Modified logit = model using proposed reject inference; Bayesian = the Bayesian model by Chen and Åstebro (2003); Bivariate = bivariate two-stage probit model; Censored = the logit model based on the censored sample; Augmentation = the Augmentation reject inference model.
2. The new credit model for 1998 NSSBF sample is different from the model based on the 1993 NSSBF model. Some independent variables are same, but many are different. Especially, the financial ratios used in the 1993 NSSBF model are not significant in the model for the 1998 NSSBF model, and they are not included in the estimation. For the sake of simplicity we do not present the details for this new credit score model, as well as the estimation for bivariate two-stage model since it is not the focus of the research.
3. The bad rate is calculated based on the sample reject rate at 17.58% (i.e., 493 observations are screened out).
4. Difference test 1 estimates differences between the Brier scores of five models (Modified logit, Bayesian, Bivariate, Censored and augmentation) and the Brier score of the Optimal model. The hypothesis is that their difference is zero. The first row shows the t values, and the second row shows the p-values.
5. Difference test 2 estimates differences between the Brier scores of four models (Bayesian, Bivariate, Censored and Augmentation) and the Brier score of the Modified logit model. The hypothesis is that their difference is zero. The first row shows the t values, and the second row shows the p-values.

Table 3: Model Performance for the 1993 NSSBF Sample.

	Optimal	*Weak selection*					*Strong selection*				
		Modified logit	Bayesian	Bivariate	Censored	Augmentation	Modified logit	Bayesian	Bivariate	Censored	Augmentation
KS	0.354	0.358	0.354	0.346	0.358	0.323	0.344	0.332	0.257	0.314	0.314
Bad rate	14.2%	14.3%	14.5%	14.3%	14.4%	14.6%	14.5%	14.6%	16.3%	15.4%	15.7%
Brier score	0.281	0.282	0.283	0.289	0.284	0.289	0.286	0.288	0.314	0.296	0.301
Difference test 1		2.967	3.421	4.96	3.824	5.093	3.968	4.738	8.999	6.518	7.62
		[0.003]	[0.001]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Difference test 2			1.766	4.952	4.69	4.616		3.166	11.248	8.493	8.048
			[0.078]	[0.000]	[0.000]	[0.000]		[0.002]	[0.000]	[0.000]	[0.000]
Logarith-mic score	0.443	0.444	0.445	0.452	0.446	0.457	0.448	0.451	0.493	0.462	0.475

1. Optimal = model based on complete sample; Modified logit = model using proposed reject inference; Bayesian = the Bayesian model by Chen and Åstebro (2003); Bivariate = bivariate two-stage probit model; Censored = the logit model based on the censored sample; Augmentation = the Augmentation reject inference model.
2. The bad rate is calculated based on the sample reject rate at 20% (i.e., 918 observations are screened out).
3. Difference test 1 estimates differences between the Brier scores of five models (Modified logit, Bayesian, Bivariate, Censored and augmentation) and the Brier score of the Optimal model. The hypothesis is that their difference is zero. The first row shows the t values, and the second row shows the p-values.
4. Difference test 2 estimates differences between the Brier scores of four models (Bayesian, Bivariate, Censored and Augmentation) and the Brier score of the Modified logit model. The hypothesis is that their difference is zero. The first row shows the t values, and the second row shows the p-values.

Table 2 and 3 compares KS, bad rate, the Brier and logarithmic scores among different models. The significances of the differences of these Brier scores are reported. The results show that the optimal model is the most efficient in both within-sample and out-of-sample tests. When compared to the optimal model, the differences of the Brier scores of the tested reject inference models are all significant[15]. The results also show that the modified logit model proposed here is better than the Bayesian reject inference method (Chen and Åstebro, 2003) in both within and out-of-sample tests. The differences are significant at 90% and 99% for weak and strong selection cases in within-sample tests. This is as expected since for a specific model this method imbeds the uncertainty of the missing data mechanism into the modeling procedure directly. On the other hand, the Bayesian approach imputes a complete sample where rejected information has been weighted. Therefore, this sample is available for other modeling while the modified logit model does not impute missing data.

The results in Table 2 and 3 provide some clear evidence that is consistent with theoretical expectations. First, from weak selection case to strong selection case the power of the proposed reject inference method decreases. This is understandable since more data loss means any method to recover lost information will be more unreliable, especially in the case of missing not at random. Secondly, when the assumption of missing not at random is strictly valid (i.e., the within-sample test), the model performance using the proposed reject inference technique is the best among all models in comparison. Especially in the case of strong selection where the bias caused by the missingness is serious, the logit model with reject inference is significantly better than the Bayesian bound and collapse approach, and also much better than the other four models statistically. It might be difficult to evaluate the economic magnitude of the differences in scores. For that purpose we calculated the bad rate in the selected sample based on the various reject inference techniques. For example, while the optimal bad rate in the 1993 sample is 14.2%, for strong selection the bad rate for the modified logit model is 14.5% while it is 14.6%, 16.3%, 15.4% and 15.7% respectively for the Bayesian model, the Heckman's model, the model based on the censored sample as well as the augmentation

[15] One exception is that in the weak selection case in the out-of-sample test the censored model is only significantly different at 90% confidence level.

model. In weak selection the magnitude of the economic effect is also visible, but not that large.

Table 2 shows that in the out-of-sample test the proposed reject inference is not better than the model based on the censored sample. The differences of the Brier scores in the weak and strong selection cases are all significant. We believe these results do not imply that reject inference does not work. The poor classification power is instead caused by underlying assumption not being satisfied – in this case we believe some data in the 1998 sample to be missing at random in addition to being missing not at random. Since, as implemented, the proposed reject inference technique cannot account fully for MAR data its classification power is reduced. There is evidence to support this claim. First, although the ideal model is still the most efficient, the difference to other models is not large, especially in the case of weak selection where the MNAR data constraint is less severe. This means that at least in the case of weak selection, all applied reject inference techniques are quite similar, and also very close to that of the ideal model. Second, the time lapsed in between the two surveys (five years) is quite long and cover different stages of the business cycle causing the models based on the 1993 and 1998 samples to be quite different. Many of the independent variables used in these two models are different. In particular, the financial ratios that are significant in the 1993 sample do not have any predictive power in the 1998 sample. This implies that the credit score based on the 1993 sample is not a strong classifier for the 1998 sample. Among all five tested models, that based on the censored sample is then most efficient as it is built on the 1998 sample. This evidence implies that in this out-of-sample test, the efficiency gained by using the censored sample overwhelms the weak bias caused by the inefficient credit screening when applying the credit scoring model based on the 1993 sample. Therefore, it is inconclusive to judge the efficiency of the proposed reject inference technique based on the 1998 data. Rather, it supports the claim that if the missingness is missing at random, a model based on the censored sample is efficient and effective.

Table 3 shows that augmentation model in the within-sample test is generally worse than the modified logit model, Bayesian model as well as model based on the censored sample, judged by the classification power. Especially, in the strong selection case of the within-sample test, this model is inferior to all other models. In the out-of–

sample test, which is possibly the case of MAR, augmentation model also does not gain anything in model improvement for classification power. These evidences show some support for the robustness of the proposed reject inference technique. It also has implications for industry which currently relies heavily on the augmentation approach. Therefore, in the case that user is not sure if the missing data mechanism is MNAR or MAR, it is likely that there is little to lose in predictive power for the credit scoring model when applying the maximum likelihood reject inference technique.

6 Conclusion

In this paper, we proposed a maximum likelihood approach for reject inference within a logit scoring model. This approach could be applied to any model based on the maximum likelihood method. We model reject inference as a missing data mechanism where we assume the missing data is non-ignorable. However, unlike the Bayesian approach (Chen and Åstebro, 2003) that uses a missing data imputation method to generate a complete sample for further modeling, this approach embeds the missing data mechanism into the modeling procedure directly.

Using the 1993 and 1998 NSSBF surveys results showed that the proposed model would improve the classification power for credit scoring when samples suffer from a selection bias caused by credit applicants being rejected. It is an efficient reject inference technique when logit regression is used for credit scoring. Similar to the Bayesian bound and collapse approach, this reject inference technique requires estimation of the missingness distribution. We demonstrate that by using the bad rate distribution information from the accepted sample (if possible, as well as the bad rate distribution from the original credit scoring model) one can effectively solve this problem. Test results confirm that the proposed reject inference technique is appropriate for the case of missing not at random. If the missing data mechanism is missing at random (or missing completely at random), the most efficient method is list-wise deletion where all observations with missing data are dropped. This analysis shows that successful implementation of reject inference depends on accurate specification of the missing data mechanism, and a proper estimation of the missingness distribution. It is a desirable direction for further research.

REFERENCES

Ash, D and Meester, S. (2002). Best practices in reject inference. Presentation at Credit Risk modeling and Decision Conference, Wharton Financial Institutions Center, Philadephia, May, 2002.

Banasik, J. B., Crook, J. N. and Thomas, L. C. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society* 54, 822 – 832.

Chen, Gongyue and Thomas Åstebro (2003). Bound and Collapse Bayesian Reject Inference When Data are Missing not at Random, in Åstebro T., P. Beling, D. Hand, B. Oliver and L. B. Thomas (Eds.) "Mathematical Approaches to Credit Risk Management," Conference Proceedings, Banff International Research Station for Mathematical Innovation and Discovery October 11-16.

Copas, J. B. and H. G. Li (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society*, B, 59, 55-95.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*, 39 (Ser. B): 1-38.

Feelders, A. J. (2000). Credit scoring and reject inference with mixture models. *International Journal of Intelligent System in Accounting, Finance and Management*, 9, 1-8.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin (1995). *Bayesian Data Analysis*. Chapman & Hall.

Graham, J. W., and Donaldson, S.I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of followup data. *Journal of Applied Psychology*, 78, 119-128.

Hand, David J. (1997). *Construction and Assessment of Classification Rules*. Chichster: Wiley.

Hand, David J. (1998). Reject inference in credit operations, in *Credit Risk Modeling: Design and Application* (ed. E. Mays), 181-190, AMACOM.

Hand, David J. (2001). Measuring Diagnostic Accuracy of Statistical Prediction Rules. *Statistica Neerlandica*, 53, 3-16.

Jacobson, Tor, and Kasper F. Roszbach (1999). Evaluating bank lending policy and consumer credit risk, in *Computational Finance 1999* (edited by Yaser S. Abu-Mostafa et al.) the MIT Press, 2000.

Lipsitz LA, Nakajima I, Gagnon M, Hirayama T, Connelly CM, Izumo H & Hirayama T (1994). Muscle strength and fall rates among residents of Japanese and American nursing homes: an international cross-cultural study. *Journal of the American Geriatrics Society* 42(9): 953–959.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Little, R.J.A (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

Paik, Myunghee Cho, Ralph Sacco, and I.-Feng Lin (2000). Bivariate binary data analysis with nonignorably missing outcomes. *Biometrics*, 56: 1145-1156.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63: 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schafer, Joseph L., and John W. Graham (2002). Missing data: our view of the state of the art. *Psychological Methods*, Vol. 7, No. 2: 147-177.

Sebastiani, Paola, and Marco Ramoni (2000). Bayesian inference with missing data using bound and collapse. *Journal of Computational and Graphical Statistics*, Vol. 9, No. 4: 779-800.

Thomas, Lyn C. (2004). A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. Working paper (a revision version), School of Management, University of Southampton, UK. (the original paper published in *International Journal of Forecasting*, Vol.16, pp. 149-172, 2000)

Appendix 1: Definitions of Variables for the 1993 NSSBF Sample

Name	Definition
Delinq	Dependent variable. = 1 if a firm had been 60 or more days delinquent on at least one business obligation within the past three years, else 0.
trad	=1 if the business operates in wholesale trade, retail trade, or services, defined by the U.S. Bureau of the Census as the following SIC 2-digit, else 0.
othliab	= 1 if the firm owned any other assets that were not listed in the survey, else 0.
payable	= 1 if the firm have any accounts payable, else 0.
ratio1	The ratio of sales 1992 over total expense
ratio3	The ratio of profit 1992 over total expense
ratio6	The ratio of sales 1992 over total liability
ratio7	The ratio of profit 1992 over total liability
small	= 1 if the firm had less than 20 equivalent full-time employees, else 0.
isminor	= 1 if more than 50 percent of the firm owned by blacks or African Americans, Asians, Pacific Islanders, American Indians, or Alaskan Natives, else 0.
young	= 1 if the age of the principal owner was smaller than or equal to 40, else 0.
workexp2	= 1 if the owner had more than 5, but smaller than and equal to 10 years work experience, else 0.
workexp3	= 1 if the owner had more than 10, but smaller than and equal to 20 years work experience, else 0.
workexp4	= 1 if the owner had more than 20 years work experience, else 0.
selfdo	= 1 if this business was founded by the current owner(s) , else 0.
newfirm	= 1 if the firm age was smaller than or equal to 5 years, else 0.
local	= 1 if the area of sales was the same area as the firm's main office, else 0.
wide	= 1 if the area of sales was notional or international, else 0.
export	= 1 if the firm exported outside of the United States, else 0.
lineloan	= 1 if during 1993 the firm had business line of credit, else 0.
lease	= 1 if during 1993 the firm had capital leases from financial institutions or other sources, else 0.
motor	= 1 if during 1993 the firm had loans on motor vehicles used primarily for business purposes, else 0.
equloan	= 1 if during 1993 the firm had loans secured by equipment, or was the firm financing any purchases of equipment by installment payments, else 0.
othloan	=1 if as of year end 1993, excluding trade credit or credit with suppliers and loans from banks, the firm have any other loans from financial institutions or from any other sources, else 0.
trade	=1 if during 1993 the firm purchase any goods or services on account during 1993 rather than pay for the purchases before or at the time of delivery, else 0.
short	=1 if during 1993 the firm had ever required financing for seasonal or unexpected short-term credit needs, else 0.
newequ	= 1 if during the last three years, the firm has obtained additional equity capital from existing owners, their relatives, or from new or existing partners, else 0.
inven	= 1 if the firm had an inventory of merchandise or production materials
isassets	= 1 if firm held any bonds; held any stocks for short-term investment; or had any prepaid expenses or other current assets, else 0.
invest	=1 if the firm was owed any money for mortgages or real estate, or did the firm have any other investments, else 0.
land	= 1 if the firm owned any land, else 0.
typeprop	= 1 if firm type is proprietorship, else 0.
typepart	=1 if firm type is partnership, else 0.
typecor	=1 if firm type is corporation, else 0.

Appendix 2: Sample Statistics for the 1993 NSSBF sample

Variable	Mean	Std. Dev.	Min	Max
deliq	0.200	0.400	0	1
trad	0.594	0.491	0	1
othliab	0.430	0.495	0	1
payable	0.748	0.434	0	1
ratio1	1.336	2.740	0	160
ratio3	0.353	2.743	-1	159
ratio6	69.041	2956.558	0	200000
ratio7	7.556	266.704	-360	18000
small	0.642	0.479	0	1
isminor	0.170	0.376	0	1
young	0.210	0.407	0	1
workexp2	0.165	0.371	0	1
workexp3	0.375	0.484	0	1
workexp4	0.383	0.486	0	1
selfdo	0.713	0.453	0	1
newfirm	0.206	0.404	0	1
local	0.538	0.499	0	1
wide	0.153	0.360	0	1
export	0.123	0.328	0	1
lineloan	0.360	0.480	0	1
lease	0.159	0.365	0	1
motor	0.270	0.444	0	1
equloan	0.501	0.500	0	1
othloan	0.143	0.350	0	1
trade	0.681	0.466	0	1
short	0.265	0.441	0	1
newequ	0.203	0.402	0	1
inven	0.623	0.485	0	1
isassets	0.280	0.449	0	1
invest	0.217	0.412	0	1
land	0.269	0.443	0	1
typeprop	0.326	0.469	0	1
typepart	0.076	0.265	0	1
typecor	0.368	0.482	0	1

Appendix 3: Credit Scoring Model Estimation using Logit Regression

Number of obs = 4589
LR chi2(33) = 527.22
Prob > chi2 = 0.0000
Log likelihood = -2033.014
Pseudo R2 = 0.1148

delinq	**Coef.**	**Std. Err.**	**z**	**P>|z|**
trad	-0.2071	0.0853	-2.429	0.015
othliab	0.2580	0.0906	2.848	0.004
payable	0.5768	0.1097	5.258	0.000
ratio1	1.2796	0.6223	2.056	0.040
ratio3	-1.4006	0.6196	-2.261	0.024
ratio6	-0.0023	0.0012	-1.928	0.054
ratio7	-0.0063	0.0035	-1.803	0.071
small	0.2061	0.1037	1.987	0.047
isminor	0.3300	0.1039	3.176	0.001
young	0.2178	0.1058	2.059	0.039
workexp2	0.3392	0.1748	1.940	0.052
workexp3	0.1645	0.1704	0.966	0.334
workexp4	0.1534	0.1838	0.835	0.404
selfdo	0.1662	0.0925	1.797	0.072
newfirm	0.2607	0.1077	2.420	0.016
local	0.0129	0.0920	0.141	0.888
wide	-0.3946	0.1385	-2.850	0.004
export	0.4923	0.1344	3.662	0.000
lineloan	-0.2273	0.0948	-2.397	0.017
lease	0.4716	0.1037	4.547	0.000
motor	0.1702	0.0882	1.930	0.054
equloan	0.1724	0.0918	1.878	0.060
othloan	0.4918	0.1042	4.719	0.000
trade	0.5396	0.0998	5.404	0.000
short	0.8761	0.0843	10.388	0.000
newequ	0.5384	0.0913	5.897	0.000
inven	0.1661	0.0911	1.824	0.068
isassets	-0.2454	0.1048	-2.342	0.019
invest	-0.2522	0.1064	-2.371	0.018
land	-0.2118	0.0982	-2.156	0.031
typeprop	0.2946	0.1215	2.424	0.015
typepart	0.1768	0.1712	1.033	0.302
typecor	0.2929	0.1061	2.760	0.006
Constant	-4.8164	0.6885	-6.996	0.000