Reject Inference in Credit Scoring

Jie-Men Mok

BMI paper
January 2009

Preface

In the Master programme of Business Mathematics and Informatics (BMI), it is required to perform research on a business related subject which involves the fundamental aspects of mathematics and computer science. This BMI paper records the findings of the literature research.

During my internship period (see *'Process Scoring for Micro Credit Loans'*, November 2008), I came across the reject inference problem. Reject inference was not included in the internship assignment due to lack of time. Therefore, this BMI paper describes the main approaches to reject inference in credit scoring.

I would like to thank my supervisor, Marianne Jonker for her guidance. Since this paper ends my Master programme of BMI, I would like to take this opportunity to thank my fellow students for the joyful years in college. These include (in alphabetic order): Aïcha, Alicia, Angelique, Chi Hung, Christel, Jade, Natasia, Safar, Suman and Ruth. Finally, special thanks go to my partner Paul Harkink for his support.

Jie-Men Mok

Amsterdam, January 2009

Summary

Reject inference is the process of estimating the risk of default for loan applicants that are rejected under the current acceptance policy. The reject inference problem is considered as a missing data problem. When the data are missing completely at random, then there is no reject inference problem at all. If the data are missing at random, then the selection mechanism is ignorable. But when the data are missing not at random, then the selection mechanism is nonignorable.

When the selection mechanism is ignorable, then logistic regression, discriminant analysis or the mixture model are suggested. In case of a nonignorable selection mechanism, Heckman's model is suggested.

In logistic regression, it is not necessary to make assumptions about the distribution of the application characteristics. Also, only the accepted applications are required. But this means that logistic regression is not able to handle the rejected applications efficiently.

Another alternative is discriminant analysis which depends on the class conditional distribution of the characteristics. But this means that the results are biased when the model is not based on the complete set of applications. The mixture model can avoid this bias by including the rejected applications in the model. Unfortunately, it is possible that the EM algorithm returns bad estimates of the parameters.

Heckman's model corrects the bivariate model by taking the omitted variables into account. But empirical studies indicate that the parameter estimators are not reliable, which may lead to unreliable estimates.

It is diffcult to compare the different approaches to reject inference, since the true creditworthiness of the rejected applications is unknown. But simulation studies may be helpful in evaluating the reject inference techniques.

Contents

List of Figures

Chapter 1

Introduction

In credit scoring, the creditworthiness of the loan applicant is assessed by means of a scoring model. The credit scoring model is based on the characteristics of the loan applicant and it estimates the credit risk by predicting the repayment behaviour of the applicant.

This chapter will first describe the loan process. Afterwards, it will be shown how a biased sample can lead to incorrect results, which is why reject inference will be introduced. Finally, the objective and the outline of this paper will be described.

1.1 Loan process

When the client applies for a loan, then the application can be accepted or rejected by the creditor. The accepted applicant will receive a loan. After a certain period of time, the loan performance of the accepted application can be labelled as good or bad. The loan process is also shown in figure 1.1.

The selection mechanism determines whether the application is accepted or rejected and the outcome mechanism determines the loan performance of the accepted application. The main objective in credit scoring is to model the outcome mechanism.

Note that the default risk of the rejected applications is unknown since the loan performance of the application can not be observed when rejected.

1.2 Sample bias

The probability that an accepted application will be a bad loan can be estimated with the given data, but the estimated probability that a rejected application is in fact a good loan might be biased. Therefore, when a new acceptance rule is based on the data on accepted applications only then this may lead to incorrect results. This will be illustrated with an example.

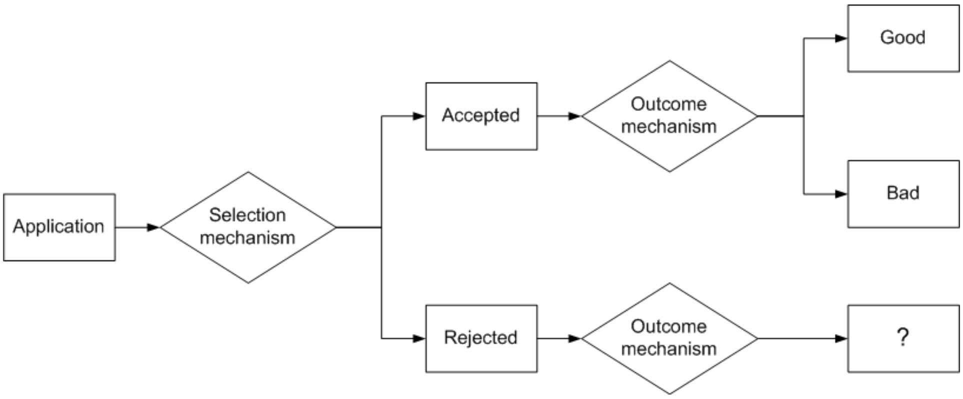

Figure 1.1: Loan process

Suppose that a creditor accepts applications randomly. When the scoring model is based on the randomly accepted applications, the results indicate that the probability that new customers will have a bad loan performance is high. Based on this result, the creditor decides to adjust their acceptance policy by creating extra restrictions for new customers.

In the new situation, the new customers will be granted a loan when they have an extremely low default risk. This results in a set of accepted applications that does not represent the total set of applications. When another scoring model is built on the accepted applications only, then it appears that the probability that new customers will have a bad loan performance is low. Without the results of the previous scoring model, this may lead to the incorrect impression that new customers have a low risk of default and the restrictions of the new customers are removed. In this case, the true risk of default will rise.

In order to obtain unbiased results, the credit scoring model should be based on randomly accepted applications. But in reality it is not feasible to obtain data on randomly accepted applications, because this increases the default risk which will result in high costs.

1.3 Reject inference

In practice, the rejected applications are included in the credit scoring model in order to avoid biased results. The process of estimating the risk of default for loan applicants that are rejected under the current acceptance policy is called reject inference. This can be considered as statistical inference with incomplete data. The reject inference methods in this paper can also be applied to other missing data problems such as insurance policy acceptance, personnel selection and medical diagnosis.

Reject inference has attracted a great deal of interest and many different methods have been proposed for the reject inference problem. But it is difficult to evaluate the performance of these methods since the true creditworthiness of the rejected applications is unknown. It is also unknown what kind of methods for reject inference are being used by banks, since they are reluctant to share this valuable information with other competitors.

1.4 Objective

This paper will present the primary models for handling reject inference in credit scoring. Each reject inference method can be applied to a certain type of data set. The models will be elaborated and compared with other models, which will finally lead to the discussion of the advantages and disadvantages of each reject inference method.

1.5 Outline of this document

The rejected applications are considered to be missing data. There are different types of missing data which will be described in chapter 2. The approach to handling reject inference depends on the type of missing data.

When the data are missing at random then the logistic regression technique, discriminant analysis and the mixture model are suggested, which are described respectively in chapters 3, 4 and 5. Chapter 6 describes Heckman's model, which can be applied in case the data are missing not at random. In the final chapter, an overview of the suggested models is given and each model is evaluated.

Chapter 2

Missing data

The data on the loan performance of the rejected applications are missing. There are three types of missing data [1]: missing completely at random, missing at random and missing not at random. The different types of missing data will be elaborated in this chapter. But first the notations will be introduced.

The characteristics of the applicant is based on what has been filled in on the application form, together with the information regarding the credit history of the applicant which can be obtained from the central credit bureau. The characteristics can be completely observed for each applicant and will be denoted as a vector of variables $\mathbf{x}= (x_1, \ldots, x_k)$.

The outcome mechanism is denoted by the class label $y \in \{0, 1\}$ and the selection mechanism by the auxiliary variable $a \in \{0, 1\}$. If the application is accepted then $a = 1$, and if it is rejected then $a = 0$. The class label y can only be observed when $a = 1$. If the accepted application has a good loan performance then $y = 1$, and if the loan performance is bad then $y = 0$.

2.1 Missing completely at random

When the applications are randomly accepted, then the class label y is missing completely at random (MCAR). This means that the probability that the application will be accepted does not depend on the characteristics of the applicant nor on the loan performance, which is formulated below.

$$\mathbb{P}(a = 1|\mathbf{x}, y) = \mathbb{P}(a = 1)$$

In that case, there is no reject inference problem since analysis of the accepted applications will yield unbiased results.

2.2 Missing at random

Many creditors use a selection model to determine which application will be accepted or not. When the selection model is based on the observable characteristics of the applicant and not on the loan performance, then the rejected applications are missing at random (MAR). In that case, the probability that the application is accepted depends on the observable characteristics of the applicant only. This is formulated as follows.

$$\mathbb{P}(a = 1|\mathbf{x}, y) = \mathbb{P}(a = 1|\mathbf{x})$$

When the equality above holds, then it can be shown that the probability of a good loan performance depends also on the observable characteristics of the applicant and not on the selection model. This is formulated below.

$$\mathbb{P}(y = 1|\mathbf{x}, a = 1) = \mathbb{P}(y = 1|\mathbf{x}, a = 0) = \mathbb{P}(y = 1|\mathbf{x})$$

That means that the loan performance of the rejected applications has the same distribution as the one of the accepted applications for any fixed value of $\mathbf{x}$.

2.3 Missing not at random

When the selection model is also based on the impression made by the applicant or other characteristics that can not be observed, then the rejected applications are missing not at random (MNAR). In that case, the probability that the application is accepted depends on the loan performance, even when conditioned on the observable characteristics of the applicant. This is formulated as follows.

$$\mathbb{P}(a = 1|\mathbf{x}, y) \neq \mathbb{P}(a = 1|\mathbf{x})$$

From the inequality above, it follows that the probability of a good loan performance depends also on the selection model when conditioned on the observable characteristics of the applicant. This is formulated below.

$$\mathbb{P}(y = 1|\mathbf{x}, a = 1) \neq \mathbb{P}(y = 1|\mathbf{x}, a = 0)$$

Therefore, the loan performance of the accepted applications has a different distibution from the one of the rejected applications for any fixed value of $\mathbf{x}$.

2.4 Overview

The selection mechanism is ignorable when the rejected applications are MAR. But if the rejected applications are MNAR, then the selection mechanism is nonignorable. In that case, the selection mechanism should be included in the model in order to obtain reliable parameter estimates of the outcome mechanism.

The following chapters will describe different approaches to handling reject inference in case the selection mechanism is either ignorable or nonignorable. An overview is given below:

- Ignorable selection mechanism (MAR)
  - Chapter 3: Logistic regression
  - Chapter 4: Discriminant analysis
  - Chapter 5: Mixture model
- Nonignorable selection mechanism (MNAR)
  - Chapter 6: Heckman's model

Chapter 3

Logistic regression

The objective in credit scoring is to model the outcome mechanism. When the data are MAR, then the selection mechanism is ignorable in the scoring model (see section 2.2). In that case, the outcome mechanism is modelled as the probability of a good loan given the characteristics of the applicant. This is denoted as follows.

$$p = \mathbb{P}(y = 1|\mathbf{x}) = 1 - \mathbb{P}(y = 0|\mathbf{x})$$

In that case, the relationship between the independent variables $\mathbf{x} = (x_1, \dots, x_k)$ and the dependent variable y can be investigated by means of logistic regression [3]. This will be further described in the following sections.

3.1 Logit function

In logistic regression, the variable y has a Bernoulli distribution with the unknown parameter p.

$$y \sim B(1, p)$$

The link between p and $\mathbf{x}$ is determined by the logit function, which models the logit of p as a linear function of $\mathbf{z}$ and θ, where $\mathbf{z} = (1, x_1, \dots, x_k)$ with weights $\theta = (\theta_0, \dots, \theta_k)^T$. This is formulated below.

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{z}\theta = \theta_0 + x_1\theta_1 + \dots + x_k\theta_k$$

When the logit function is rewritten, then p can be defined as follows.

$$p(\theta) = \frac{1}{1 + e^{-\mathbf{z}\theta}} \tag{3.1}$$

Figure 3.1 shows an example of logistic regression with a one-dimensional $\mathbf{x}$, where the datapoints with (x, y)-coordinates are estimated by the logistic curve.

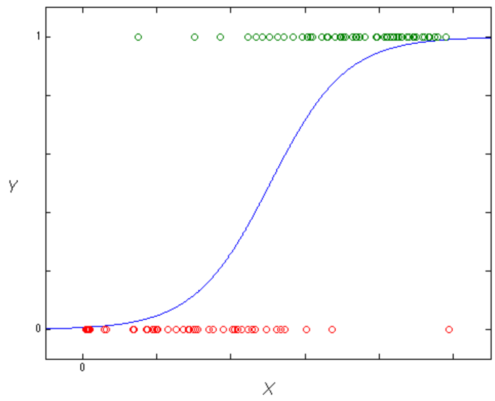

Figure 3.1: Logistic regression

3.2 Maximum likelihood estimation

The vector of parameters θ can be estimated by the maximum likelihood estimator (MLE) $\hat{\theta}$. The MLE estimates the parameters by maximizing the loglikelihood function. The loglikelihood function $ll_b(\theta)$ for y given $\mathbf{x}$ is formulated as follows.

$$ll_b(\theta) = \sum_{i=1}^{N} y_i \ln(p(\theta)) + (1 - y_i) \ln(1 - p(\theta))$$

When taking the derivative of $ll_b(\theta)$ with respect to θ, then the MLE $\hat{\theta}$ can be found by taking the θ that sets the derivative to zero. This requires an iterative procedure, like the Newton-Raphson method or Fisher's scoring method.

3.3 Cut-off score

The class label y of an incoming application is based on the estimation of the probability p. The estimated probability $\hat{p}$ can be obtained by plugging the characteristics of the applicant $\mathbf{x}$ and the parameter estimates $\hat{\theta}$ in (3.1). Given a cut-off score c, when $\hat{p} \geq c$ then the application is labelled as a good loan but if $\hat{p} < c$ then the application is labelled as a bad loan. The cut-off score depends on the risk that the creditor is willing to take.

There are two types of errors in statistics: Type I error and Type II error. In credit scoring, the Type I error occurs when an actual bad loan is labelled as a good loan, and if a good loan is labelled as a bad loan then this is considered as a Type II error. The cut-off score depends on the Type I error which is allowed by the creditor.

Chapter 4

Discriminant analysis

The logistic regression technique directly estimates the probability of a good loan given the application characteristics. But in discriminant analysis, the probability is indirectly estimated by means of Bayes theorem. In both cases, the data are assumed to be MAR (see section 2.2).

In the following sections, the discriminant analysis will be further elaborated [4]. At the end of the chapter, it will be shown that discriminant analysis leads to biased results.

4.1 Bayes theorem

In discriminant analysis, the conditional probability of a good loan is modelled by means of Bayes theorem below with class conditional probability function $p_j(\mathbf{x}) = \mathbb{P}(\mathbf{x}|y = j)$ and unconditional probability $\pi_j = \mathbb{P}(y = j)$ for discrete $\mathbf{x}$ and class $j \in \{0, 1\}$.

$$\mathbb{P}(y = 1|\mathbf{x}) = \frac{\pi_1 p_1(\mathbf{x})}{\pi_0 p_0(\mathbf{x}) + \pi_1 p_1(\mathbf{x})} \tag{4.1}$$

When the data sample is divided into two subsets with the same class label $y = j$, then each subset can be used separately to estimate π_j and $p_j(\mathbf{x})$ in order to obtain an estimate of $\mathbb{P}(y = 1|\mathbf{x})$. The following sections show how the prior probability π_j and posterior probability $p_j(\mathbf{x})$ can be modelled as a discriminant function.

4.2 Decision rule

The objective of discriminant analysis is to divide the vector space of $\mathbf{x}$ into regions $\Omega = \{\Omega_0, \Omega_1\}$ where $\mathbf{x}$ lies in Ω_j if y is classified as j, such that the probability of making a classification error is minimized with the Bayes minimum error rule. This is formulated below where two scenarios for '>'

and '<' are denoted by '$\gtrless$' which leads to classification of respectively Ω_0 and Ω_1.

$$\mathbb{P}(y=0|\mathbf{x}) \gtrless \mathbb{P}(y=1|\mathbf{x}) \Rightarrow \mathbf{x} \in \left\{ \begin{array}{l} \Omega_0 \\ \Omega_1 \end{array} \right.$$

When using Bayes theorem in (4.1), then the decision rule above can be rewritten in the following likelihood ratio form.

$$\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \gtrless \frac{\pi_1}{\pi_0} \Rightarrow \mathbf{x} \in \left\{ \begin{array}{l} \Omega_0 \\ \Omega_1 \end{array} \right.$$

If a bad loan is misclassified as a good loan, then this is a Type I error. The probability of a Type I error is denoted by α, which is derived as follows.

$$\begin{aligned} \alpha &= \mathbb{P}(y=0, \mathbf{x} \in \Omega_1) \\ &= \int_{\Omega_1} \mathbb{P}(\mathbf{x}, y=0)\mathrm{d}\mathbf{x} \\ &= \int_{\Omega_1} \pi_0 p_0(\mathbf{x})\mathrm{d}\mathbf{x} \end{aligned}$$

The Type II error occurs if a good loan is misclassified as a bad loan. The probability of a Type II error is denoted by β, which is derived in the same way as α. See below.

$$\beta = \int_{\Omega_0} \pi_1 p_1(\mathbf{x})\mathrm{d}\mathbf{x}$$

When the goal is to minimize the probability of the Type II error where the probability of the Type I error is set to a fixed level, then this can be formulated as an optimization problem where the objective is to minimize β subject to a fixed α. The minimum can be found by means of the Lagrange function Λ_β with Lagrange multiplier λ_β.

$$\Lambda_\beta = \int_{\Omega_0} \pi_1 p_1(\mathbf{x})\mathrm{d}\mathbf{x} + \lambda_\beta \left(\int_{\Omega_1} \pi_0 p_0(\mathbf{x})\mathrm{d}\mathbf{x} - \alpha \right) \tag{4.2}$$

Note that the whole space of $\mathbf{x}$ is $\Omega_0 \cup \Omega_1$ which means that the following equation holds.

$$\int_{\Omega_0} \pi_1 p_1(\mathbf{x})\mathrm{d}\mathbf{x} = 1 - \int_{\Omega_1} \pi_1 p_1(\mathbf{x})\mathrm{d}\mathbf{x}$$

When the first term on the right-hand side of (4.2) is substituted by the right-hand side of the equation above, then this leads to the following result.

$$\Lambda_\beta = 1 - \lambda_\beta \alpha + \int_{\Omega_1} \Big(\lambda_\beta \pi_0 p_0(\mathbf{x}) - \pi_1 p_1(\mathbf{x}) \Big)\mathrm{d}\mathbf{x}$$

The Lagrange function above will be minimized when Ω_1 is chosen such that the following holds.

$$\lambda_\beta \pi_0 p_0(\mathbf{x}) < \pi_1 p_1(\mathbf{x}) \Rightarrow \mathbf{x} \in \Omega_1$$

When the decision rule above is rewritten, then this leads to the following likelihood ratio.

$$\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \gtrless \frac{\pi_1}{\lambda_\beta \pi_0} \Rightarrow \mathbf{x} \in \left\{ \begin{array}{l} \Omega_0 \\ \Omega_1 \end{array} \right.$$

For specified π_j and $p_j(\mathbf{x})$, the Lagrange multiplier λ_β can be estimated by means of numerical methods.

4.3 Linear discriminant function

The decision rules in the previous section have been expressed as functions of $\mathbf{x}$ via $p_j(\mathbf{x})$, where only the relative magnitude matters and not the absolute values of $p_j(\mathbf{x})$. Therefore, the decision rule can be formulated more generally with discriminant function $g(\mathbf{x})$ and a constant d.

$$g(\mathbf{x}) \gtrless d \Rightarrow \mathbf{x} \in \left\{ \begin{array}{l} \Omega_0 \\ \Omega_1 \end{array} \right.$$

The discriminant function can have the form of any monotonic function without effecting the decision rule.

When the discriminant function is assumed to be linear, then this will have a low analytical complexity with computational advantages. In that case, the decision rule and the linear discriminant function $h(\mathbf{x})$ with threshold weight w_0 and weight vector w can be formulated as follows.

$$h(\mathbf{x}) = w_0 + w^T\mathbf{x}$$

$$h(\mathbf{x}) \gtrless 0 \Rightarrow \mathbf{x} \in \left\{ \begin{array}{l} \Omega_0 \\ \Omega_1 \end{array} \right.$$

The weights in the linear discriminant function can be estimated by means of the following approaches: linear programming formulation, error correction approaches, Fisher's method and least squares methods.

An example of linear discriminant analysis is shown in figure 4.1 with a two-dimensional $\mathbf{x}$, where the datapoints are separated by the linear discriminant line $h(\mathbf{x}) = 0$. The upper area is classified as Ω_1 and the lower area as Ω_0, where the green dots represent good loans and the red dots bad loans. When a datapoint lies on the discriminant line, then it can be assigned to either class or be left undefined.

Figure 4.2 shows an example of the distributions $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ which are represented by respectively the red and green curves, where two scenarios are shown for $C1$ and $C2$ which are the lines used to discriminate between good loans and bad loans. In case of $C1$, the red and purple areas represent the Type I error. When $C2$ is used, then the green and purple areas represent the Type II error.

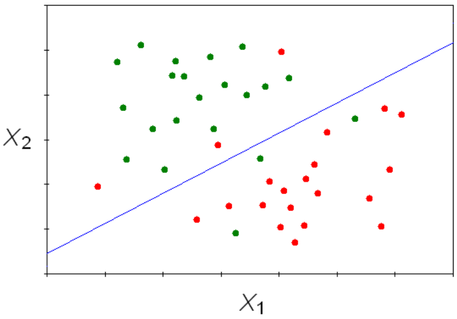

Figure 4.1: Linear discriminant analysis

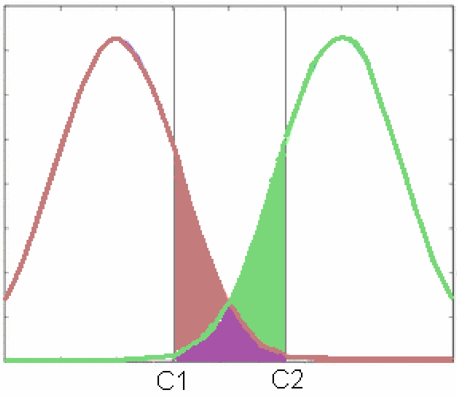

Figure 4.2: Type I and Type II errors

4.4 Bias in distribution

When the scoring model depends on the distribution $p_j(\mathbf{x})$ for $j \in \{0,1\}$ that are estimated from data which are missing at random, then this leads to incorrect classification of the loan performance. This will be illustrated with an example below.

In figure 4.3, the distribution $p(\mathbf{x}|y = j)$ for the complete set of applications and the truncated distribution $p(\mathbf{x}|a = 1, y = j)$ for the accepted applications only are shown in respectively the upper and lower figure [2]. In the complete set of applications, the good loans have a higher mean than the bad loans.

Assume that the cut-off score is located between the means of the good loans and bad loans. In that case, many expected bad loans are rejected which strongly affects the distribution $p(\mathbf{x}|a = 1, y = 0)$. As a result, the mean of the bad loans is overestimated and the variance is underestimated. The distribution $p(\mathbf{x}|a = 1, y = 1)$ on the other hand, is hardly affected.

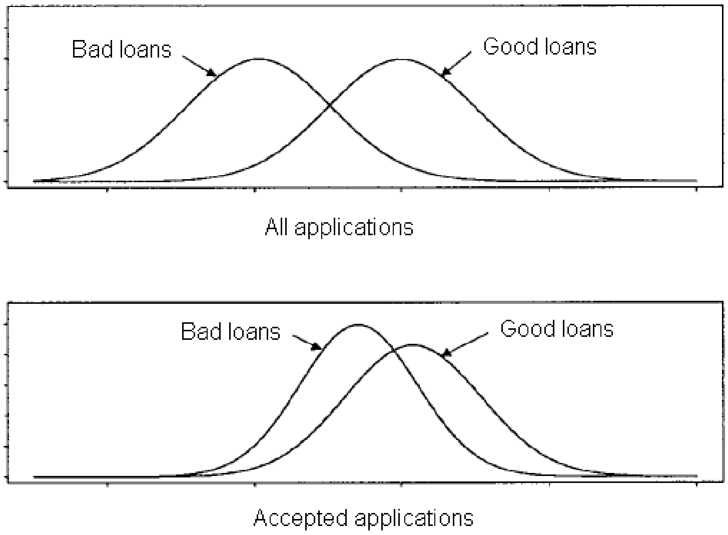

Figure 4.3: Bias in distribution of characteristics

The discrimininant function is based on the class conditional distribution of the characteristics. This means that the estimates of $p_j(\mathbf{x})$ for $j \in \{0,1\}$ that are based on the accepted applications only, are biased due to the acceptance rule. The magnitude of the bias depends on the true distribution and the acceptance rule.

The bias in the distribution can be avoided by including the characteristics of the rejected applications in the estimation process of the conditional distributions. This will be described in the next chapter.

Chapter 5

Mixture model

This chapter will describe the mixture model of the probability distribution [8], where the data are assumed to be MAR (see section 2.2). The mixture model includes the rejected applications in the estimated distribution of the characteristics, which avoids the bias that resulted in discriminant analysis.

5.1 Two-component mixture distribution

When the applications consist of good loans and bad loans, then the probability distribution of $\mathbf{x}$ can be modelled as a finite mixture distribution with two components. See figure 5.1 for an example where the red and green curves represent respectively the distributions $p(\mathbf{x}, y = 0)$ for the bad loans and $p(\mathbf{x}, y = 1)$ for the good loans, and the blue curve represents the mixture distribution $p(\mathbf{x})$ for the complete set of loans.

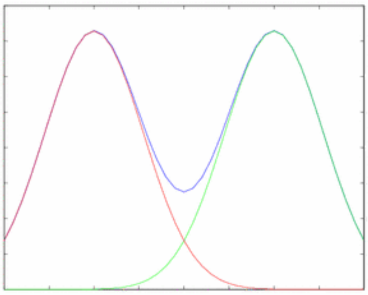

Figure 5.1: Two-component mixture distribution

In the mixture distribution, it is necessary to make assumptions about the parametric density of $\mathbf{x}$. The probability function $p(\mathbf{x})$ is derived as follows, where the prior probability π_j is considered as the proportion of

$p_j(\mathbf{x}, \vartheta_j)$ in the mixture model with unknown parameter ϑ_j for $j \in \{0, 1\}$.

$$\begin{aligned} p(\mathbf{x}) &= p(\mathbf{x}, y = 0) + p(\mathbf{x}, y = 1) \\ &= \pi_0 p_0(\mathbf{x}, \vartheta_0) + \pi_1 p_1(\mathbf{x}, \vartheta_1) \end{aligned}$$

Note that $\pi_0 + \pi_1 = 1$ where $0 \leq \pi_j \leq 1$.

In reality, the outcome of the accepted applications only are observed but not the rejected applications. When taking the missing data problem into account, then the likelihood function l_i with $\mathbf{x}_i$ and y_i for observation i is defined as follows.

$$l_i = \begin{cases} \pi_0 p_0(\mathbf{x}_i, \vartheta_0) + \pi_1 p_1(\mathbf{x}_i, \vartheta_1) & \text{if } y_i \text{ is missing} \\ \pi_j p_j(\mathbf{x}_i, \vartheta_j) & \text{if } y_i = j \text{ for } j \in \{0, 1\} \end{cases}$$

When there are m rejected applications and n accepted applications, then the loglikelihood function of the incomplete data ll_{inc} is formulated as follows.

$$\begin{aligned} ll_{\text{inc}} &= \sum_{i=1}^{m} \log\{\pi_0 p_0(\mathbf{x}_i, \vartheta_0) + \pi_1 p_1(\mathbf{x}_i, \vartheta_1)\} \\ &\quad + \sum_{i=m+1}^{m+n} (1 - y_i) \log\{\pi_0 p_0(\mathbf{x}_i, \vartheta_0)\} + y_i \log\{\pi_1 p_1(\mathbf{x}_i, \vartheta_1)\} \end{aligned}$$

The parameter vector $\varphi = (\pi_0, \pi_1, \vartheta_0, \vartheta_1)$ can be estimated by the maximum likelihood estimator $\hat{\varphi}$. The MLE is the set of parameters which maximizes the loglikelihood function. The loglikelihood function $ll_{\text{inc}}(\varphi)$ is a complex function of φ. In order to find the MLE of φ, this requires a special computational algorithm.

5.2 EM algorithm

The expectation-maximization (EM) algorithm is an iterative procedure, which can be used to compute the MLE when the data is incomplete. The maximum of the incomplete-data loglikelihood function is estimated by optimizing the complete-data loglikelihood function ll_{com}, which is less complex than $ll_{\text{inc}}(\varphi)$. The complete-data loglikelihood function ll_{com} is defined as follows.

$$ll_{\text{com}} = \sum_{i=1}^{m+n} (1 - y_i) \log\{\pi_0 p_0(\mathbf{x}_i, \vartheta_0)\} + y_i \log\{\pi_1 p_1(\mathbf{x}_i, \vartheta_1)\}$$

The objective of the EM algorithm is to maximize ll_{com} by means of iterative optimization of the expectation of ll_{com}.

The EM algorithm has two steps: expectation step (E-step) and maximization step (M-step). This will be further described in the following subsections.

5.2.1 E-step

The EM algorithm starts with arbitrary parameter estimates $\varphi^{(0)}$. Since the likelihood function may have multiple (local) maxima, the initial values of the parameter estimates are critical for finding the global maximum. Therefore, the procedure should be repeated with different starting values because a poor choice of initial values may lead to bad estimates.

In the E-step, the expectation of ll_{com} is formulated by replacing y_i with $\mathbb{E}[y|\mathbf{x}_i]$. Note that $\mathbb{E}[y|\mathbf{x}_i] = \mathbb{P}(y = 1|\mathbf{x}_i)$. When $\varphi^{(t-1)}$ is given in iteration t, then the estimate of $\mathbb{P}(y = j|\mathbf{x}_i)$ is denoted by ϱ_j for $j \in \{0, 1\}$ which can be calculated by means of Bayes theorem in (4.1) as follows.

$$\varrho_j(\mathbf{x}_i, \varphi^{(t-1)}) = \frac{\pi_j^{(t-1)} p_j(\mathbf{x}_i, \vartheta_j^{(t-1)})}{\pi_0^{(t-1)} p_0(\mathbf{x}_i, \vartheta_0^{(t-1)}) + \pi_1^{(t-1)} p_1(\mathbf{x}_i, \vartheta_1^{(t-1)})}$$

The expectation of ll_{com} is denoted by $Q(\varphi|\varphi^{(t-1)})$ for iteration t which is formulated as follows.

$$Q(\varphi|\varphi^{(t-1)}) = \sum_{i=1}^{m+n} \varrho_0(\mathbf{x}_i, \varphi^{(t-1)}) \log\{\pi_0 p_0(\mathbf{x}_i, \vartheta_0)\} + \varrho_1(\mathbf{x}_i, \varphi^{(t-1)}) \log\{\pi_1 p_1(\mathbf{x}_i, \vartheta_1)\}$$

5.2.2 M-step

In the M-step, a new set of parameters $\varphi^{(t)}$ is computed such that $Q(\varphi|\varphi^{(t-1)})$ is maximized for iteration t. This is formulated below, where the computation of $\pi_s^{(t)}$ and $\vartheta_s^{(t)}$ for $s \in \{0, 1\}$ will be described in this subsection.

$$\varphi^{(t)} = \arg\max_{\varphi} Q(\varphi|\varphi^{(t-1)})$$

The prior probability $\pi_s^{(t)}$ for $s \in \{0, 1\}$ can be derived by means of the Lagrange function Λ_Q with Lagrange multiplier λ_Q, where the objective is to maximize Q subject to the constraint $\pi_0 + \pi_1 = 1$. See below.

$$\Lambda_Q = Q(\varphi|\varphi^{(t-1)}) + \lambda_Q \left(\sum_{i=0}^{1} \pi_j - 1 \right)$$

When the partial derivatives of Λ_Q are set to zero, then this results in the following set of equations.

$$\frac{\partial \Lambda_Q}{\partial \pi_s} = \lambda_Q + \frac{1}{\pi_s} \sum_{i=1}^{m+n} \varrho_s(\mathbf{x}_i, \varphi^{(t-1)}) = 0 \qquad (5.1)$$

$$\frac{\partial \Lambda_Q}{\partial \lambda_Q} = \sum_{j=0}^{1} \pi_j - 1 = 0 \qquad (5.2)$$

When π_s in (5.1) is solved and substituted in (5.2), then this leads to the following result.

$$\sum_{j=0}^{1}\left(-\frac{1}{\lambda_Q}\sum_{i=1}^{m+n}\varrho_s(\mathbf{x}_i,\varphi^{(t-1)})\right)=1\Longrightarrow-\sum_{i=1}^{m+n}\sum_{j=0}^{1}\varrho_s(\mathbf{x}_i,\varphi^{(t-1)})=\lambda_Q$$

Note that $\varrho_0(\mathbf{x}_i,\varphi^{(t-1)})+\varrho_1(\mathbf{x}_i,\varphi^{(t-1)})=1$. When this is plugged in the formula above, then this leads to the following result.

$$-(m+n)=\lambda_Q$$

When the result above is substituted in (5.1), then $\pi_s^{(t)}$ for iteration t can be formulated as follows.

$$\pi_s^{(t)}=\frac{1}{m+n}\sum_{i=1}^{m+n}\varrho_s(\mathbf{x}_i,\varphi^{(t-1)})$$

The component parameter $\vartheta_s^{(t)}$ for $s\in\{0,1\}$ is computed such that the relevant part of $Q(\varphi|\varphi^{(t-1)})$ is maximized, see below.

$$\vartheta_s^{(t)}=\arg\max_{\vartheta_s}\sum_{i=1}^{m+n}\varrho_s(\mathbf{x}_i,\varphi^{(t-1)})\log\{\pi_s p_s(\mathbf{x}_i,\vartheta_s)\}$$

The computation of $\vartheta_s^{(t)}$ can be further specified when the parametric density is given. For some distributions, it is possible to obtain an analytical expression for $\vartheta_s^{(t)}$.

The E- and M-steps are repeated until the stopping criteria with convergence level $\xi>0$ is met, where $Q(\varphi^{(t)})$ and $Q(\varphi^{(t-1)})$ result from respectively iteration $t+1$ and iteration t.

$$|Q(\varphi^{(t)})-Q(\varphi^{(t-1)})|<\xi$$

When π_j and $p_j(\mathbf{x},\vartheta_j)$ are finally estimated, then $\mathbb{P}(y=j|\mathbf{x})$ can be computed by means of Bayes theorem in (4.1).

Chapter 6

Heckman's model

In the previous three chapters, different reject inference methods have been proposed where data are assumed to be MAR (see section 2.2). When data are MNAR (see section 2.3), then the selection mechanism is nonignorable. In that case, the performance of the rejected applications can be inferred with Heckman's model [6] which will be described in this chapter.

6.1 Bivariate distribution

In Heckman's model, the selection mechanism a_i and the outcome mechanism y_i are modelled by respectively the unobserved numeric variables a_i^* and y_i^* for observation i.

$$a_i = \begin{cases} 0 & \text{if application is rejected:} \quad a_i^* < 0 \\ 1 & \text{if application is accepted:} \quad a_i^* \geq 0 \end{cases}$$

$$y_i = \begin{cases} 0 & \text{if loan is bad:} \quad y_i^* < 0 \\ 1 & \text{if loan is good:} \quad y_i^* \geq 0 \end{cases}$$

Note that the loan performance y_i is only observed when the application is accepted ($a_i = 1$). For a complete dataset, the variables a_i^* and y_i^* are defined as follows with parameters β and γ and random noise d_i and e_i .

$$\begin{aligned} a_i^* &= \mathbf{x}_i\beta + d_i \\ y_i^* &= \mathbf{x}_i\gamma + e_i \end{aligned}$$

The random errors d_i and e_i are assumed to be bivariate normally distributed with mean μ as vector of zeros and variance-covariance matrix Σ where ρ is the unknown correlation between d_i and e_i.

$$\begin{bmatrix} d_i \\ e_i \end{bmatrix} \sim N(\mu, \Sigma) \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

6.2 Omitted variables

The reject inference problem is a sample selection bias, which is caused by omitting variables in the bivariate model. This will be illustrated by means of the regression function.

In case of a complete dataset, the regression function for the outcome mechanism can be formulated as follows for observation i.

$$\mathbb{E}[y_i^*] = \mathbf{x}_i\gamma$$

When data are available on accepted applications only, then the regression function is formulated as follows.

$$\begin{aligned}\mathbb{E}[y_i^*|a_i^* \geq 0] &= \mathbf{x}_i\gamma + \mathbb{E}[e_i|a_i^* \geq 0] \\ &= \mathbf{x}_i\gamma + \mathbb{E}[e_i|d_i \geq -\mathbf{x}_i\beta]\end{aligned}$$

If the parameter γ is estimated by omitting the final term in the equation above, then the reject inference problem arises.

The sample selection bias can be corrected by formulating the conditional expectation as follows with correlation ρ and hazard function H_i.

$$\mathbb{E}[e_i|d_i \geq -\mathbf{x}_i\beta] = \rho H_i$$

The hazard function is also called the inverse Mills ratio, and it consists of density function ϕ and cumulative distribution function Φ for the standard normal distribution.

$$H_i = \frac{\phi(-\mathbf{x}_i\beta)}{1 - \Phi(-\mathbf{x}_i\beta)} \tag{6.1}$$

When the omitted variables are included, then the outcome mechanism is modelled as follows.

$$\mathbb{E}[y_i^*|a_i^* \geq 0] = \mathbf{x}_i\gamma + \rho H_i \tag{6.2}$$

6.3 Two-step estimation

The parameters in Heckman's model can be estimated in two steps. In the first step, the parameter β is estimated by probit analysis in order to obtain an estimate of H_i. When H_i is estimated, then the parameters γ and ρ are estimated by least squares estimation in the second step. This will be further described in the following subsections.

6.3.1 Probit analysis

The objective in the first step is to estimate H_i, which is done by estimating β. The MLE of β can be computed by means of probit analysis as follows, where the data is assumed to be complete.

$$\mathbb{E}[a_i] = \Phi(\mathbf{x}_i\beta)$$

The corresponding loglikelihood function ll_p is derived as follows.

$$\begin{aligned} ll_p(\beta) &= \sum_{i=1}^{N} a_i \ln \mathbb{P}(a_i = 1) \\ &= \sum_{i=1}^{N} a_i \ln \mathbb{E}(a_i) \\ &= \sum_{i=1}^{N} a_i \ln \Phi(\mathbf{x}_i \beta) \end{aligned}$$

The MLE of β is obtained by taking β which sets the derivative of ll_p to zero. When the MLE of β is plugged in (6.1), then this leads to the estimate $\hat{H}_i$.

6.3.2 Least squares estimation

After obtaining the estimate $\hat{H}_i$, the resulting model can be considered as the ordinary least squares problem where the parameters γ and ρ in (6.2) are estimated by the least squares estimators (LSE). The LSEs are the γ and ρ that minimize the sum of squared errors (SSE). The SSE is defined as follows.

$$SSE = \sum_{i=1}^{N} (y_i - \mathbb{E}[y_i^* | a_i^* \geq 0])^2$$

In case of no sample bias, the errors d_i and e_i are independent. Therefore, the presence of sample bias can be indicated by testing the null hypothesis that $\rho = 0$. The hypothesis can be tested by means of the Wald test, likelihood ratio test or Lagrange multiplier test.

6.4 Robustness

Empirical research indicates that the estimators of Heckman's model are not robust [7]. It is assumed that the bivariate model is linear with errors that are normal distributed and homoscedastic. When the assumptions do not hold, then the estimates are not reliable.

It also appears that collinearity problems between the explanatory variables often arise. In literature, extensions of Heckman's model or alternative approaches are suggested which can yield more reliable parameter estimates.

Chapter 7

Conclusion

An important question is whether reject inference can really improve the performance of the credit scoring model [5]. It is difficult to compare the different reject inference techniques, since the true creditworthiness of the rejected applications is unknown. That explains why little empirical studies have been published about the comparison of reject inference techniques. Simulation studies may be helpful in evaluating the reject inference techniques.

This chapter will first give an overview of the proposed models and then each model is evaluated.

7.1 Overview

This paper describes the main approaches to reject inference. The choice of the reject inference method depends on the type of missing data. There are three types of missing data: missing completely at random, missing at random and missing not at random. When data are MCAR, then there is no reject inference problem at all. In case of MAR, the selection mechanism is ignorable. But if data are MNAR, then the selection mechanism is nonignorable.

This paper describes four reject inference techniques which can be applied in the following situations:

- Ignorable selection mechanism (MAR)
  - Logistic regression
  - Discriminant analyisis
  - Mixture model
- Nonignorable selection mechanism (MNAR)
  - Heckman's model

7.2 Evaluation

For an ignorable selection mechanism, both logistic regression and discriminant analysis perform reject inference by extrapolating the outcome of the rejected applications from the outcome of the accepted applications. This means that when the fraction of accepted applications is relatively small, then the foundation of the scoring model may be weak due to bad extrapolation.

In logistic regression, no assumptions are made about the class conditional distribution of the application characteristics. Therefore, logistic regression is able to perform the extrapolation unbiased when it is based on the accepted applications only. On the other hand, the performance of the model will not be improved when the rejected applications are included. This means that when the characteristics of the rejected applications are available, then logistic regression is not able to handle the available information efficiently.

Unlike logistic regression, discriminant analysis depends on the class conditional distribution of the characteristics, which means that the results are biased when the model is not based on the complete set of applications. The mixture model can avoid this bias by including the characteristics of the rejected applications in the model. Unfortunately, it is possible that the EM algorithm returns a local maximum which leads to bad estimates of the mixture model. Note that in the mixture model, it is necessary to make assumptions about the parametric density of the characteristics.

For a nonignorable selection mechanism, Heckman's model corrects the bivariate model by taking the omitted variables into account. But empirical studies indicate that the parameter estimators are not reliable, which may lead to unreliable estimates. In literature, extensions or alternatives are suggested which can produce more reliable estimates.

Bibliography

[1] Feelders, A.J. (2003), *'An overview of model based reject inference for credit scoring'*. Technical report, Utrecht University, Institute for Information and Computing Sciences.
http://www.cs.uu.nl/people/ad/mbrejinf.pdf

[2] Feelders, A.J. (1999), *'Credit scoring and reject inference with mixture models'*. International Journal of Intelligent Systems in Accounting, Finance and Management, 8:271-279.

[3] Gunst, M (2005), *'Statistical models'*. Lecture notes, vrije Universiteit, Amsterdam.

[4] Hand, D.J. (1981), *'Discrimination and classification'*, John Wiley & Sons, New York.

[5] Hand, D.J. and Henley, W.E. (1993), *'Can reject inference ever work?'*, IMA Journal of Mathematics Applied in Business and Industry, 5(4):45-55.

[6] Heckman, J. J. (1979), *'Sample selection bias as a specification error'*, Econometrica, 47, l, 53-161.

[7] Puhani, P. A. (2000), *'The Heckman correction for sample selection and its critique'*, Journal of Economic Surveys, 14, 1.

[8] Rouse, D.M. (2005), *'Estimation of Finite Mixture Models'*. Thesis, North Carolina State University.
http://www.lib.ncsu.edu/theses/available/etd-11282005-140114/unrestricted/etd.pdf