Bayesian Logistic Regression Models for Credit Scoring

by

Gregg Webster

A thesis submitted to Rhodes University in partial fulfilment of the requirements for the degree of

Master of Commerce

in

Mathematical Statistics

December 2011

Supervisor: Professor S.E. Radloff

DECLARATION

Except for references specifically indicated in the text, this study is my own work and has not been submitted elsewhere for degree purposes.

_____

Gregg Webster

Abstract

The Bayesian approach to logistic regression modelling for credit scoring is useful when there are data quantity issues. Data quantity issues might occur when a bank is opening in a new location or there is change in the scoring procedure. Making use of prior information (available from the coefficients estimated on other data sets, or expert knowledge about the coefficients) a Bayesian approach is proposed to improve the credit scoring models. To achieve this, a data set is split into two sets, "old" data and "new" data. Priors are obtained from a model fitted on the "old" data. This model is assumed to be a scoring model used by a financial institution in the current location. The financial institution is then assumed to expand into a new economic location where there is limited data. The priors from the model on the "old" data are then combined in a Bayesian model with the "new" data to obtain a model which represents all the available information. The predictive performance of this Bayesian model is compared to a model which does not make use of any prior information. It is found that the use of relevant prior information improves the predictive performance when the size of the "new" data is small. As the size of the "new" data increases, the importance of including prior information decreases.

Acknowledgements

Thank you to my parents, Robert and Debra Webster for their constant love and support throughout my studies. This research is dedicated to them.

Thank you to my supervisor, Professor Sarah Radloff for her guidance throughout my research. Your helpfulness and support is highly appreciated.

Thank you to my friends and family for their support and input. Another thank you goes out to all the staff and students of the Rhodes Statistics department. A particular thank you goes to Tafadzwa Mutengwa for all the help, laughs and hard work we did together.

This research was possible through the assistance from the Henderson Rhodes University Prestigious Scholarship and is hereby acknowledged.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Context of the Research

Consumer credit is one of the main driving forces which allowed for the rise (and possible demise) of most of the leading industrialized countries. The growth in home ownership and consumer spending over the last 50 years would not have occurred without credit. When a financial institution grants credit to an applicant the financial institution trusts the applicant to pay back the credit. The applicant may, however, default on payments back to the institution. It is the task of the financial institution to make sure that the number of defaults is minimized so that risk is reduced. This is done by screening the applicants when they apply for credit. Scoring methods are used to estimate the credit worthiness of an applicant. These credit scoring methods estimate the probability that an applicant will default or become delinquent. Credit scoring methods use statistical methods based on historical credit data to build a model which predicts whether an applicant will default or not. The financial institution can then use the model to decide whether or not to grant credit to the applicant also considering how much risk the institution is willing to take on.

As mentioned, building a credit scoring model requires the use of historical data. There may, however, be situations when there is limited historical data. This might occur when the financial institution is expanding into a new economic location (country) and no data is available at first. Data quantity issues might also occur when there is a change in the scoring procedure. In these situations it is difficult to build a good scoring model as there is initially not enough data available. Thus, expert information can be important. An existing reliable generic scoring model may be available at first which could be used for scoring. This generic scoring model could then be modified as new data becomes available. Institutions already using scorecards may be able to combine their expert knowledge with new sources of information to obtain improved scoring models. In order to do this, a Bayesian approach is proposed where the expert knowledge is combined with the limited amount of data. The aim is to see whether the combination of expert knowledge with data gives a better model than one that uses only the limited amount of data.

The scope of Bayesian inference has greatly improved since it was discovered that Markov Chain Monte Carlo (MCMC) Methods could be used to sample from the posterior distributions. The general MCMC algorithm is called the Metropolis-Hastings (MH) algorithm.

1.2 Objectives of the Study

The objectives of this study are as follows:

- Investigate credit scoring and the associated problems - such as reject inference.
- Introduce the concepts and methods of the Bayesian logistic regression models for credit scoring. This includes an in-depth explanation of the Markov Chain Monte Carlo (MCMC) methods.
- Develop a standard logistic regression scorecard.
- Develop a Bayesian approach to the scorecard for when the bank enters a new market or there is a change in procedure.
- Compare the Bayesian approach to the standard logistic regression approach. This would involve comparing the models' predictive powers on a test set.
- Make recommendations on the Bayesian approach to credit scoring.

1.3 Organization of the Study

Chapter 2 gives the history of credit scoring, problems with credit scoring and examines previous research on models used for credit scoring. The chapter provides a literature review on the models used for credit scoring focusing on the Bayesian logistic regression models. Chapter 3 examines the methods used in detail; it provides derivations and proofs of key results in order to gain an understanding of the models used. In Chapter 4 the results of the data analyses are presented and discussed. Chapter 5 summarizes the study, gives limitations and discusses areas for further research.

Chapter 2: Literature Review

2.1 History of Credit Scoring

Credit scoring is essentially a classification problem where applicants are classified into different groups. According to Thomas (2009) statistical classification techniques started when Fisher (1936) developed one of the first successful classification models to classify three different types of the iris flower. He used different physical measurements of the flower to discriminate between the three types of Iris flowers. Durand (1941) was then the first to recognise that these statistical classification techniques could be used to classify good and bad loans. Before this, Thomas (2009) states that financial institutions based decisions on whether to grant credit subjectively. When credit cards were introduced in the 1960s, the usefulness of credit scoring started to be realized. Because of the large number of people applying for credit cards, automation of the credit application procedure seemed to be the only solution. When the financial institution introduced the credit scoring model they found that the model performed a lot better than the previous (subjective) judgment scheme. The result was that, as Thomas (2009) states, default rates dropped by 50% or more. In the 1980s the success of credit scoring in credit cards meant that financial institutions started using scoring methods for other products too such as personal loans, home loans and business loans.

The subprime mortgage crisis caused a global recession in 2007. This crisis proved that financial institutions did not fully understand the risks they were taking on. According to Rona-Tas and Hiß (2008) a credit score generally used by financial institutions in the U.S.A. is the Fair Isaac Co. (FICO®) score. They state that these FICO scores grew steadily from 2000 to 2005. This made subprime borrowers appear less risky. Possible reasons for these inflated FICO scores include the data used to construct the FICO scores are historical data, not necessarily only from subprime lenders, and banks putting pressure on credit rating agencies to inflate their credit rating scores. The reason why banks would put pressure on credit rating agencies is that they were able to sell their loans to investors. Thus, the banks would want to grant as many loans as possible and then sell them to investors.

2.2 Overview of Credit Scoring and Credit Scoring Methods

Because credit scoring is fundamentally a classification problem, there are a number of methods available for credit scoring. Hand and Henley (1997) give a review in statistical classification methods in consumer credit scoring. They first give an overview of credit scoring and building a scoring model including some associated problems. They mention that scorecards are classifiers which "use predictor variables from application forms and other sources to yield estimates of the probabilities of defaulting" (Hand and Henley, 1997, p. 524). A threshold on this probability is then obtained, classification applied and a decision on whether a loan should be granted or not, can be given on a new applicant. They further explain that when building a credit scoring model, three approaches to selecting the variables are commonly used, as follows:

- Using expert knowledge. Where an experienced industry expert decides what variables will fit the data well;
- Using stepwise statistical methods such as forward/backward stepwise methods which sequentially add/delete variables;
- Selecting individual variables by using a measure of difference between the distributions of the good and bad risks on that variable.

A major problem in credit scoring is that of reject inference. Mok (2009) explains that complete data are only available for accepted applicants. This means that the observed behaviour of an applicant is only available for the accepted applicants. Because the accepted applicants were already accepted through an existing scoring model, we have biased data. It would be better to build a model where everyone is accepted and their behaviour is observed. However, this is unfeasible for banks. Therefore to solve this bias problem, reject inference is proposed. According to Mok (2009) this is "the process of estimating the risk of default for loan applicants that are rejected under the current acceptance policy" (Mok, 2009, p. 1). Crook and Banasik (2002) suggest finding a cut-off to classify the rejects whether good or bad then include these rejected applicants in the new model.

Hand and Henley (1997) give an overview of different models used for credit scoring. These methods are discriminant analysis, regression analysis, logistic regression, probit

analysis, mathematical programming, recursive partitioning (decision trees), expert systems, neural networks, nonparametric smoothing methods and time varying models. They state that "there is no overall best model" (Hand and Henley, 1997, p. 535). This is because the best model depends on the data structure. It is also mentioned that neural networks might provide a good modelling approach when there is poor understanding of the data structure. However, these models provide a "black box" approach and usually no understanding can be gained from the model.

There have been a number of studies which compare these methods in credit scoring. Altman *et al*. (1994) provided one of the first investigations of neural networks in credit scoring. Neural networks were compared to linear discriminant analysis (LDA) and it was found that LDA performed better. Desai *et al*. (1996) obtained different results. Using a credit union data set, a neural network performed better than LDA but did not perform significantly better than logistic regression. In a master's degree study by Komorád (2002), logistic regression is compared to multilayer perceptron and radial basis function neural networks for credit scoring. These models were trained and their performance tested on confidential data from a French bank. It was found that the multilayer perceptron neural network and the radial basis function neural network gave very similar results but the logistic regression performed the best.

Thomas (2009) claims that logistic regression is the most commonly used method for the construction of scorecards. Logistic regression is part of a wider class of generalized linear models (GLMs) as shown by Nelder and Wedderburn (1972). The reason for this is that the binomial distribution, which is the distribution of the response in logistic regression, is part of the exponential family of distributions. GLMs include a number of models such as normal linear regression, logistic regression, Poisson regression etc. One of the first applications of logistic regression to credit scoring is given by Steenackers and Goovaerts (1989). Based on data from a Belgian credit company they develop a logistic regression model. Nineteen predictor variables were utilized and then using stepwise logistic regression, 11 variables were chosen for a final model. Steenackers and Goovaerts (1989) also mentioned that the model relies on historical data. Therefore, a periodical review of the model is necessary to adjust for shifts in the underlying factors. To solve this problem in credit scoring, Whittacker *et al*. (2007) developed a Kalman filter for a credit scorecard. Here, the scorecard is updated by combining the new applicant data with the previous best estimate. A Bayesian approach can also be used to update a model - the posterior

distribution is updated as soon as new information becomes available. Greenberg (2008) stated that Bayesian updating is a very attractive feature of Bayesian inference. With Bayesian logistic regression, numerical methods are used to update the model. The reason for this is that conjugate priors (the posterior distribution comes from the same family of the prior distribution) do not exist. A popular method used to update the model is the Markov Chain Monte Carlo (MCMC) method.

2.3 Markov Chain Monte Carlo Methods

Conjugate priors for the logistic regression model do not exist which makes sampling from the posterior distribution difficult. Gelfand and Smith (1990) introduced the turning point for the use of MCMC methods in statistics. These MCMC methods are methods which are used to obtain samples from a posterior distribution when it is not analytically possible to obtain the posterior. MCMC methods were introduced by statistical physicists in the 1950s. Metropolis *et al*. (1953) introduced an algorithm known as the Metropolis algorithm. The algorithm was then generalized by Hastings (1970) and became the Metropolis-Hastings (MH) algorithm. The algorithm works by constructing a Markov chain which has a stationary distribution equal to the target (posterior) distribution. This is achieved through a kind of accept-reject strategy. A value is proposed and this value is accepted or rejected according to a rule which ensures that the Markov chain generated has a stationary distribution equal to the target (posterior) distribution. It resulted in renewed interest in Bayesian statistics through the use of modern computers being able to perform algorithms - such as the Gibbs sampler and Metropolis-Hastings (MH) algorithm. Determining integrals is of vital importance in obtaining the posterior distribution. The Metropolis-Hastings algorithm is more general than the Gibbs sampler. The MH algorithm is the principle algorithm on which Bayesian logistic regression is based. The MH algorithm adopts a kind of accept-reject strategy to the simulation while the Gibbs sampler is a special case, which can be used when it is possible to sample from conditional distributions. Most studies which consider Bayesian logistic regression use the MH algorithm to sample from the posterior (Ziemba, 2005; Wilhelmsen *et al*., 2009). This is because the Gibbs sampler cannot be used directly as one cannot sample easily from the

conditional distributions. Holmes and Held (2006), however, demonstrated how inference can be done efficiently in the Bayesian logistic regression model using a Gibbs sampler. They showed that the conditional likelihood of the regression coefficients is multivariate normal when certain auxiliary variables are introduced. This, then, allows for efficient simulation using a block Gibbs sampler. Simulation of the posterior distribution is thus either done using the Gibbs sampler or the MH algorithm. The MH algorithm is, however, far more popular with Bayesian logistic regression as the model is not complicated by additional auxiliary variables.

2.4 Studies on Bayesian Logistic Regression for Credit Scoring

There have been a number of papers which use a Bayesian approach to credit risk modelling. Mira and Tenconi (2004) developed a Bayesian hierarchical logistic regression model to predict credit risk of companies which fall in different sectors. They used fairly vague priors for the parameters of the model - priors centred at zero with large variances. They used MCMC methods to estimate the model. One method was the delayed rejection (DR) strategy with a single delaying step. This is similar to the MH algorithm but there is another chance to accept a move. Here, upon rejection of a move, a second stage candidate is proposed and accepted with a probability that preserves the so-called detailed balance condition. It is claimed that the DR estimates have a smaller variance than the estimates obtained via MH. The DR strategy has a shorter run time than the standard MH algorithm. This is the principle advantage of DR. Mira and Tenconi (2004) show how simulation using the delayed rejection strategy outperforms the standard MH algorithm in terms of efficiency of the estimates. They also show, using cross validation, that the Bayesian model outperforms the classical logistic regression model.

In another study, Ziemba (2005) showed how a (existing) generic scoring model can be updated using Bayesian methods. He mentions that this is a preferred solution in the banking industry when an international bank is opening a branch in a new country, a financial institution starts offering new services or a bank is offering services to a new group of customers. Therefore, unlike Mira and Tenconi (2004) where a fairly vague prior was used, Ziemba (2005) uses an existing model as a source of prior information for the

model parameters. He assumes that these prior parameters are normally distributed. Ziemba (2005) considers a case where a new procedure is introduced to the credit scoring - customers were required to complete an extended application form resulting in an increase in the number of predictor variables. The parameters of the model used before the change in procedure were used as priors for the parameters in the new model. For the additional variables under the new procedure, vague priors were used. The model was then updated as new data became available. Like Mira and Tenconi (2004) the Metropolis-Hastings algorithm is used to obtain the posterior but the DR was not investigated. Results are given for different amounts of new data. It was found that, when the amount of new data is smaller, including prior information results in much better accuracy than when the amount of new data is larger. The rate of this accuracy decreases as the amount of new data increases and prior information becomes less relevant.

In a similar study, Löffler *et al*. (2005) proposed a Bayesian method for banks to improve their credit scoring models by imposing prior information. This methodology enables banks with small data sets to improve their default probability estimates by making use of prior information. This might occur when a bank introduces a new rating system or expands into a new market as Ziemba (2005) mentions. Löffler *et al*. (2005) set up a simulation study in order to investigate the Bayesian approach. They bootstrapped from an initial small data set. A large data set was simulated and this was labelled "external" data. Prior information for regression coefficients were obtained from these data by running a logistic regression. A smaller data set was then simulated and named "internal" data. A logistic regression was run on this "internal" data, as well as a Bayesian logistic regression using the parameters from the "external" data as priors. This approach is very similar to Ziemba (2005) where a generic scorecard is updated. Here, the model from the "external" data can be seen as a generic scorecard. Löffler *et al*. (2005) found that when there is no structural difference between the "internal" and "external" data the Bayesian logistic regression model performs significantly better. In a more realistic case, there will be some structural differences between the "internal" and "external" data. They imposed structural differences by assuming that some variables are missing in the "external" or prior data set. It was found that the Bayesian logistic regression model still performs better than the logistic regression model when there are structural differences. Like Ziemba (2005) it was found that as the size of the "internal" data increases the relevance of prior information decreases.

In a different study, Wilhelmsen *et al*. (2009) compared the method of Integrated Nested Laplace Approximation (INLA) to MCMC methods for Bayesian modelling of credit risk. The MCMC method they used is the MH algorithm. Therefore, like Mira and Tenconi (2004) this is a comparative study between two methods to sample from the posterior. INLA can be used as an alternative to MCMC methods. They used the Bayesian formulation of logistic regression. Like Ziemba (2005) normal priors were used for the regression coefficients. INLA only allows the use of normal priors. They gave an outline of how priors for the regression coefficients can be obtained from prior information on the default probabilities. They suggested that a beta distribution for the default probability should be assumed. Greenberg (2008) stated that the beta distribution is a good choice for a prior since it is defined on the relevant range and it can produce a wide variety of shapes. Data from a Norwegian bank were used to compare INLA to MCMC when a vague and specific prior is used. They found that INLA and MCMC gave approximately the same posterior results for their particular data set, but mentioned that results may differ in other situations. They also indicated that there may be convergence issues with MCMC.

In a recent study, Fernandes *et al.* (2011) compare some different models to calculate probability of default in a low default setting. A data set consisting of a portfolio of low defaulting companies in Brazil was considered. There were 1,327 companies in the data set of which 50 defaulted. Four techniques were used to analyse the data, classical logistic regression, Bayesian logistic regression, limited logistic regression and an artificial oversampling technique. For the Bayesian logistic regression model, a non-informative prior was used. The prior was assumed to be normally distributed with zero mean and very large variance. A Gibbs sampler was used to solve the MCMC algorithm, however, the details of how this was done was not given. The four modelling procedures were compared using the area under the Response Operating Characteristic (ROC) curve, Gini coefficient and Kolmogorov-Smirnov statistics. The results showed that the four models considered gave very similar parameter estimates. However, after a bootstrap simulation was run to minimise the problem of the low number of defaults in the sample, the results revealed that the Bayesian model presented a high level of performance with a lower bootstrap variance. The Bayesian logistic regression model was, therefore, considered as the best model in this situation.

Chapter 3: Methodology and Theoretical Considerations

3.1 Methodology

A credit scoring data set analysed by Wielenga, Lucas and Georges (1999) was obtained. This is a home equity data set and the aim is to predict whether an applicant will eventually default or be seriously delinquent on a loan that allows owners to borrow against the equity of their homes. The data set consists of loan performance for 5,960 home equity loans. The dependent variable is a dummy variable indicating whether a default occurred during the duration of the loan. The proportion of applicants who defaulted in the data set is approximately 20%. There are twelve independent variables. These variables are: the reason for obtaining the credit, the type of job the applicant has, the amount of the loan request, the amount due on the existing mortgage, the value of the current property, the applicants debt-to-income ratio, the number of years the applicant has been working at a current job, the number of major derogatory reports, the number of trade lines (this is the number of other loans the applicant currently has), the number of delinquent trade lines, the age of the oldest trade line and the number of recent credit inquiries.

It is assumed that the bank is expanding into a new economic location or there is a change in procedure. The goal is to produce a good scoring model in the new location or under the new procedure when there are limited data available. Expert knowledge from the current location or under the old procedure is to be incorporated into the model at the new location or under the new procedure. It is assumed that there is a change in the economic location. The scoring procedure in the current economic location is assumed to be exactly the same as in the new economic location. Therefore, exactly the same variables are used to model good and bad applicants. To replicate this situation, the home equity data set is split as follows:

- 50% of the observations are randomly selected and labelled as the set of observations that are “old”. These observations are assumed to come from the current or home economic location.

- 10% of the observations are randomly selected and labelled as the set of observations that are “new”. These observations are assumed to come from the new or foreign economic location.
- 10% of the observations are randomly selected and used as a validation set from which, an optimal cut-off probability will be obtained. These observations are assumed to come from the current or home economic location.
- The remaining randomly selected 30% of the observations are used as test data. The “old” data set is used as prior information and the “new” data for the new procedure. These observations are assumed to come from the new economic location and are used to assess the performance of the models which are fitted on the limited amount of data in the new economic location.

To ensure that each random selection has a proportion of approximately 20% bad applicants, a stratified random sampling procedure is used.

The following steps are then undertaken:

- The data set is first checked and cleaned. This means removing outliers and estimating missing values etc.
- A logistic regression model is fitted to the “old” data set. The coefficients here are used as prior information when the “new” procedure is either introduced or the business expanded into a new market.
- An optimal cut-off probability is obtained on the validation data using the model fitted on the “old” data.
- A logistic regression model is fitted to the “new” data.
- A Bayesian logistic regression model is fitted to the “new” data using the coefficients from the “old” data set as priors.
- A Bayesian logistic regression model with non-informative prior is fitted to the “new” data.
- The performances of the logistic regression model and the Bayesian logistic regression model fitted on the “new” data are compared on the test data.
- The performances of the models are also considered using different sizes of the “new” data.

3.2 Bayesian Statistics

3.2.1 Bayesian inference

Bayesian inference provides a useful way to combine expert knowledge (prior belief) with data to arrive at some posterior belief. All Bayesian inference is conducted through the use of Bayes' theorem (Press, 1989; Bernardo and Smith, 2000; Lee, 2004; Greenberg, 2008; Ntzoufras, 2009).

Press (1989) explains that when one has a prior belief (called a prior distribution) before one observes the data, Bayes' theorem gives a mathematical procedure for updating the prior belief to arrive at a posterior distribution. The derivation of Bayes' theorem makes use of conditional probabilities,

$P(A|B) = P(A \cap B)/P(B)$ and $P(B|A) = P(B \cap A)/P(A)$.

Therefore, $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

which leads to Bayes' theorem: $P(A|B) = [P(B|A)P(A)]/P(B)$. (3.1)

3.2.2 Prior density, likelihood and posterior density functions

Following Greenberg (2008) and setting $A = \theta$ (a parameter or vector of parameters) and $B = y$, we have the following for continuous or general y.

$$\pi(\theta|y) = f(y|\theta)\pi(\theta)/f(y) \quad (3.2)$$

where $f(y) = \int f(y|\theta)\pi(\theta)d\theta$. Equation (3.2) is the basis of Bayesian statistics and econometrics. We now analyse Equation (3.2) in detail. $\pi(\theta|y)$, the left-hand-side of Equation (3.2) is the posterior density function of $\theta \mid y$. $f(y|\theta)$ is the density function of the observed data y when the parameter value is θ. $f(y|\theta)$ is called the likelihood function and is a function of θ once the data are known. $\pi(\theta)$ is called the prior density and

represents beliefs about the distribution of θ before seeing the data y. These beliefs can come from the researcher's knowledge or from other external sources. The prior distribution usually depends on parameters called hyperparameters. $f(y)$ normalizes the posterior distribution so that integrating Equation (3.2) with respect to θ yields 1. Equation (3.2) can also be written as

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta). \quad (3.3)$$

The right-hand-side of Equation (3.3) does not integrate to 1 but it has the same shape as $\pi(\theta|y)$.

The posterior distribution contains all the information we have about θ.

Bayesian updating

Equation (3.3) can be seen as a way of updating information. Our prior knowledge is updated with data. Then, as new data become available the posterior distribution is updated using Bayes' theorem. Greenberg (2008) explains this process: let θ be the parameter (or a vector of parameters) of interest and y_1 be the first set of data available. We have,

$$\pi(\theta|y_1) \propto f(y_1|\theta)\pi(\theta). \quad (3.4)$$

Now, suppose a new data set y_2 is obtained and we want the posterior distribution given all the available data. Thus,

$$\pi(\theta|y_1, y_2) \propto f(y_1, y_2|\theta)\pi(\theta) = f(y_2|y_1, \theta)f(y_1|\theta)\pi(\theta) \quad \text{using Equation (3.1)}$$

$$\propto f(y_2|y_1, \theta)\pi(\theta|y_1), \quad \text{because } \pi(\theta|y_1) \propto f(y_1|\theta)\pi(\theta)$$

from Equation (3.4).

If the data sets y_1 and y_2 are independent $f(y_2|y_1, \theta)$ simplifies to $f(y_2|\theta)$. We, therefore, obtain

$$\pi(\theta|y_1, y_2) \propto f(y_2|\theta)\pi(\theta|y_1). \quad (3.5)$$

From Equation (3.5) we can see that the posterior distribution in Equation (3.4) is now the prior distribution in Equation (3.5). Ntzoufras (2009) shows how Equation (3.5) can be generalized for a number of different data sets

$$\pi(\theta|y_1, \dots, y_t) \propto f(y_t|\theta) \dots f(y_1|\theta)\pi(\theta)$$
$$= \prod_{k=1}^{t} f(y_k|\theta)\pi(\theta).$$

Thus, as new information becomes available, the posterior distribution becomes the prior distribution for the next experiment.

Large samples

It is important to examine how the posterior distribution behaves in large samples. When there are independent trials, the likelihood function is $L(\theta|y) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} L(\theta|y_i)$. The log-likelihood function is then

$$l(\theta|y) = \log L(\theta|y)$$
$$= \sum_{i=1}^{n} l(\theta|y_i)$$
$$= n\bar{l}(\theta|y)$$

where $\bar{l}(\theta|y) = (\frac{1}{n}) \sum l(\theta|y_i)$ is the mean log-likelihood contribution (Greenberg, 2008). The posterior distribution can now be written as

$$\pi(\theta|y) \propto \pi(\theta)L(\theta|y)$$
$$\propto \pi(\theta) \exp\left(n\, \bar{l}(\theta|y)\right). \quad (3.6)$$

Now, from Equation (3.6), we see that the posterior distribution is proportional to the product of the prior distribution and an exponential term raised to the power n times a number. Thus, for large n, the exponential term dominates $\pi(\theta)$ which does not depend on n. Therefore, the larger the sample size, the less role the prior distribution will play in the posterior distribution (Greenberg, 2008).

3.2.3 Prior distributions

Specification of the prior distribution is important in Bayesian inference because it influences the posterior inference (Ntzoufras, 2009). In literature, often a prior with a normal distribution is used. The prior mean and variance is very important for specification of the prior. Ntzoufras (2009) explains that the prior mean provides a prior point estimate for the parameter of interest, while the variance gives an indication of the uncertainty on this estimate. A strong prior belief corresponds to a small prior variance and visa versa. When there is no prior information available, a prior is specified that will not influence the posterior distribution. Such a distribution is called a non-informative or vague prior distribution. Non-informative priors are often improper prior distributions in the sense that they are not integrable i.e. their integral is infinite. One can use improper priors as long as the resulting posterior is proper (Ntzoufras, 2009).

Conjugate priors

Ntzoufras (2009) states that the posterior distribution is often not analytically tractable. This can be solved by using conjugate prior distributions. This allows integrals involved in the problem to be solved analytically. A conjugate prior distribution has the property of resulting to a posterior of the same distributional family. Lee (2004) provides a definition. Let L be a likelihood function $L(\theta|y)$. A class Ω of prior distributions is said to form a conjugate family if the posterior density $\pi(\theta|y) \propto \pi(\theta)L(\theta|y)$ is in the class Ω for all y whenever the prior density is in Ω.

Training sample priors

Greenberg (2008) notes that when you have very little information on which to base a prior distribution, it is possible to train priors, providing you have a large number of observations. The idea is to make use of Bayesian updating. Greenberg gives the following

method: “a portion of the sample is selected as the training sample. It is combined with a relatively non-informative prior (a prior with a large variance and a mean of zero) to yield a first-stage posterior distribution” (Greenberg, 2008, p 53). This is then used as the prior for the remainder of the sample.

3.3 Generalized Linear Models

3.3.1 Introduction

Generalized linear models (GLMs) were introduced by Nelder and Wedderburn (1972). These models are an extension to the normal linear regression models and are based on the exponential family of distributions. A GLM has the basic structure $g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ where $\mu_i = E(Y_i)$, g is a smooth monotonic “link function”, X_i is the ith row of a model matrix, $\boldsymbol{X}$, and $\boldsymbol{\beta}$ is a vector of unknown parameters. Also, Y_i belongs to some exponential family distribution. The exponential family includes many distributions such as the Poisson, Binomial, Gamma, Normal and Inverse Gaussian distributions. A distribution belongs to the exponential family of distributions if its probability density function has the form

$$f_\theta(y) = \exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)] \tag{3.7}$$

where a, b and c are arbitrary functions, ϕ is an arbitrary dispersion parameter which represents the scale, and θ is known as the canonical parameter, which represents location.

The expectation and variance of Y are now derived. The log-likelihood of θ given a particular y is

$$\ln[f_\theta(y)] = l(\theta) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi).$$

Differentiating with respect to θ gives $\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$. Therefore, $E\left(\frac{\partial l}{\partial \theta}\right) = \frac{E(Y) - b'(\theta)}{a(\phi)}$.

Using the result that $E\left(\frac{\partial l}{\partial \theta}\right) = 0$, the expectation of Y is

$$E(Y) = b'(\theta). \tag{3.8}$$

Now, finding the second derivative with respect to θ gives $\frac{\partial^2 l}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)}$.

Using the general result $E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E(\frac{\partial l}{\partial \theta})^2$, we have

$$\frac{-b''(\theta)}{a(\phi)} = -E(\frac{Y - b'(\theta)}{a(\phi)})^2 .$$

Hence $\frac{b''(\theta)}{a(\phi)} = \frac{E(Y-b'(\theta))^2}{(a(\phi))^2}$, which leads to the variance for Y

$$\text{var}(Y) = b''(\theta)a(\phi) . \tag{3.9}$$

If ϕ is known, there is no difficulty working with GLMs using any function of $a(\phi)$. If, however, ϕ is unknown, it is common practice to assume $a(\phi) = \phi/w$, where w is a known constant. Hence, $\text{var}(Y) = b''(\theta)\phi/w$. (3.10)

Since the binomial distribution will be used in this study, it is important to show that the binomial distribution is a member of the exponential family. The probability mass function of a binomial distribution is

$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$ for $y = 0, 1, 2, \dots, n$, where p is the probability of success.

We have

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$= \exp(\ln[\binom{n}{y} p^y (1-p)^{n-y}])$$

$$= \exp\left(\ln\binom{n}{y} + y\ln(p) + (n-y)\ln(1-p)\right)$$

$$= \exp\left(\ln\binom{n}{y} + y\ln(p) + n\ln(1-p) - y\ln(1-p)\right)$$

$$= \exp\left(\ln\binom{n}{y} + y[\ln(p) - \ln(1-p)] + n\ln(1-p)\right)$$

$$= \exp\left(\ln\binom{n}{y} + y\ln\left(\frac{p}{1-p}\right) + n\ln(1-p) \right)$$

$$= \exp\left[\frac{y\ln\left(\frac{p}{1-p}\right) + n\ln(1-p)}{1} + \ln\binom{n}{y} \right]. \quad (3.11)$$

Comparing Equation (3.7) to Equation (3.11), we see that

$\theta = \ln(\frac{p}{1-p})$, $a(\phi) = 1$, $b(\theta) = -n\ln(1-p)$, and $c(y,\phi) = \ln\binom{n}{y}$. Therefore, the binomial distribution is a member of the exponential family. $\theta = \ln(\frac{p}{1-p})$ is the canonical link function and is called the logit link. The canonical link is mathematically and computationally convenient. However, other choices may also be used. The parameters of a GLM can be estimated using maximum likelihood and an iterative procedure called Iteratively Re-weighted Least Squares (IRWLS).

3.3.2 Maximum likelihood estimation

A GLM has the basic structure $g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ where $\mu_i = E(Y_i)$, $Y_i \sim f_{\theta_i}(y_i)$ and $f_{\theta_i}(y_i)$ indicates an exponential family distribution. Since the Y_i are mutually independent, the likelihood of $\boldsymbol{\beta}$ is

$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i)$. Thus the log-likelihood of **β** is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln[\, f_{\theta_i}(y_i)]$$

$$= \sum_{i=1}^{n} [\frac{y_i\theta_i - b_i(\theta_i)}{a_i(\phi)} + c_i(\phi, y_i)]$$

where ϕ is assumed to be the same for all *i*. For practical purposes, it is reasonable to assume that $a_i(\phi) = \phi/w_i$, where w_i is a constant. Therefore,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\frac{w_i}{\phi}[\, y_i\theta_i - b_i(\theta_i)] + c_i(\phi, y_i)].$$

Differentiating with respect to β_j gives

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{w_i}{\phi} [y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i) \frac{\partial \theta_i}{\partial \beta_j}].$$

Using the chain rule,

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

and Equation (2.2), we have $E(Y_i) = b_i'(\theta_i) = \mu_i$.

Hence, $\frac{\partial \mu_i}{\partial \theta_i} = b_i''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b_i''(\theta_i)}$, which leads to

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{w_i}{\phi} \left[\frac{y_i}{b_i''(\theta_i)} \frac{\partial \mu_i}{\partial \beta_j} - \frac{b_i'(\theta_i)}{b_i''(\theta_i)} \frac{\partial \mu_i}{\partial \beta_j} \right]$$

$$= \sum_{i=1}^{n} \frac{w_i}{\phi} \left[\frac{y_i - \mu_i}{b_i''(\theta_i)} \right] \frac{\partial \mu_i}{\partial \beta_j}.$$

Now, from Equation (2.3) and using the assumption $a_i(\phi) = \phi / w_i$, we obtain

$V(\mu_i) = b_i''(\theta_i)\phi / w_i$. Hence

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{[y_i - \mu_i]}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}$$

which implies that the equations to solve for $\boldsymbol{\beta}$ are given by

$$\sum_{i=1}^{n} \frac{[y_i - \mu_i]}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \text{ for all } j.$$

These are the equations that need to be solved for non-linear weighted least squares, if the weights $V(\mu_i)$ are known in advance and are independent of $\boldsymbol{\beta}$. In this case the least squares objective is, therefore

$$S = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)} \qquad (3.12)$$

where, μ_i depends non-linearly on $\boldsymbol{\beta}$ but the weights, $V(\mu_i)$ are fixed (Wood, 2006).

An iterative procedure is needed to solve the Equations (3.12). Let $\widehat{\boldsymbol{\beta}}^{[k]}$ denote the estimated parameter vector at the k^{th} iteration. Also let $\boldsymbol{\eta}^{[k]}$ be the vector with elements $\eta_i^{[k]} = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}^{[k]}$ and $\boldsymbol{\mu}^{[k]}$ be the vector with elements $\mu_i^{[k]} = g^{-1}(\eta_i^{[k]})$, where $g^{-1}(.)$ is the inverse function of the link function. Define a diagonal matrix $\boldsymbol{V}_{[k]}$ where $V_{[k]ii} = V(\mu_i^{[k]})$, then Equation (3.12) becomes

$$S = \left\| \sqrt{\boldsymbol{V}_{[k]}^{-1}}[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})] \right\|^2.$$

Replacing $\boldsymbol{\mu}$ with its first order Taylor expansion around $\widehat{\boldsymbol{\beta}}^{[k]}$ gives

$$S \approx \left\| \sqrt{\boldsymbol{V}_{[k]}^{-1}}[\boldsymbol{y} - \boldsymbol{\mu}^{[k]} - \boldsymbol{J}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{[k]})] \right\|^2$$

where $\boldsymbol{J}$ is the Jacobian matrix with elements, $J_{ij} = \frac{\partial \mu_i}{\partial \beta_j}|_{\widehat{\boldsymbol{\beta}}^{[k]}}$. Now

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} \Rightarrow g'(\mu_i)\frac{\partial \mu_i}{\partial \beta_j} = X_{ij}.$$

Thus,

$$J_{ij} = \frac{X_{ij}}{g'\left(\mu_i^{[k]}\right)}.$$

Therefore, defining a diagonal matrix $\boldsymbol{G}$ with elements $G_{ii} = g'\left(\mu_i^{[k]}\right)$, we have

$\boldsymbol{J} = \boldsymbol{G}^{-1}\boldsymbol{X}$. Hence, we obtain

$$\begin{aligned} S &\approx \left\| \sqrt{\boldsymbol{V}_{[k]}^{-1}}[\boldsymbol{y} - \boldsymbol{\mu}^{[k]} - \boldsymbol{G}^{-1}\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{[k]})] \right\|^2 \\ &= \left\| \sqrt{\boldsymbol{V}_{[k]}^{-1}}\boldsymbol{G}^{-1}[\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}^{[k]}) - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{[k]}] \right\|^2 \\ &= \left\| \sqrt{\boldsymbol{W}^{[k]}}[\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}^{[k]}) + \boldsymbol{\eta}^{[k]} - \boldsymbol{X}\boldsymbol{\beta}] \right\|^2 \\ &= \left\| \sqrt{\boldsymbol{W}^{[k]}}[\boldsymbol{z}^{[k]} - \boldsymbol{X}\boldsymbol{\beta}] \right\|^2 \end{aligned}$$

where by definition of pseudo data, $z_i^{[k]} = g'(\mu^{[k]})\left(y_i - \mu_i^{[k]}\right) + \eta_i^{[k]}$, and the diagonal weight matrix, $\boldsymbol{W}^{[k]}$, has elements $W_{ii}^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu^{[k]})^2}$ (Wood, 2006).

The following procedure is then iterated until convergence:

1. Using the current $\boldsymbol{\mu}^{[k]}$ and $\boldsymbol{\eta}^{[k]}$ obtain the pseudo data $\boldsymbol{z}^{[k]}$ and the iterative weights $\sqrt{\boldsymbol{W}^{[k]}}$.
2. Minimize the sum of squares $\left\|\sqrt{\boldsymbol{W}^{[k]}}[\boldsymbol{z}^{[k]} - \boldsymbol{X\beta}]\right\|^2$ with respect to $\boldsymbol{\beta}$ in order to obtain $\widehat{\boldsymbol{\beta}}^{[k+1]}$, and hence $\boldsymbol{\eta}^{[k+1]} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{[k+1]}$ and $\boldsymbol{\mu}^{[k+1]}$.
3. Set k to $k + 1$ and repeat until $\widehat{\boldsymbol{\beta}}$ converges.

It is common practice to use as initial values $\mu_i^{[0]} = y_i$ and $\eta_i^{[0]} = g\left(\mu_i^{[0]}\right)$ or a small adjustment to $\mu_i^{[0]}$ if $y_i = 0$.

3.3.3 Diagnostics

Model diagnostics can be divided into two types: checking (1) for outliers and influential observations and (2) the assumptions of the model.

Residual plots are very useful plots to check the adequacy of the model. For Generalized Linear Models (GLMs) the Pearson and deviance residuals (Faraway, 2006) usually provide good plots to look at because they are comparable to the standardized residuals used for the linear models. In our case, however, the outcome variable is binary which means that the plots have limited use.

However, one can consider influential observations and outliers. Multi-collinearity amongst the independent variables can also be considered.

According to Faraway (2006), for the linear model, $\widehat{\boldsymbol{y}} = \boldsymbol{Hy}$, where $\boldsymbol{H}$ is the hat matrix that projects the observed data onto the fitted values, the diagonal elements of $\boldsymbol{H}$ are the leverages h_i and represent the potential of the point to influence the fit of the model. For GLMs (and thus logistic regression) leverages are different. The IRWLS algorithm used to

fit the GLM makes use of weights, w. These weights affect the leverage. With $\boldsymbol{X}_{n\times p}$ and matrix $\boldsymbol{W} = diag(w)$, the hat matrix is

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}.$$

The diagonals of $\boldsymbol{H}$ are the leverages h_i. A large leverage value h_i indicates that the fit may be sensitive to the response at case i. Leverage measures the potential to affect the fit of the model.

Measures of influence assess the effect of each case on the fit of the model (Faraway, 2006). Influential points can be examined by looking at the Cook's distance statistic:

$$D_i = \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^T (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\hat{\phi}}$$

where the dispersion parameter ϕ is equal to 1 when the distribution is binomial (Equation 3.11). The way these leverage and Cook's distance statistics are checked is by considering their half-normal plots. Faraway (2006) explains that for a GLM, we do not expect the residuals to be normally distributed and, therefore, it is better to use half-normal plots to identify outliers. Here sorted values are compared to values of the quantiles of the half-normal distribution:

$$\Phi^{-1}\left(\frac{n+i}{2n+1}\right) \quad \text{for } i = 1, \ldots, n.$$

We then look for outliers which may be identified as points off the trend.

If some predictors are linear combinations of others, then X^TX is singular. When this happens there are serious problems with the estimation of the parameters. Collinearity amongst the predictor variables can be detected in various ways:

1. Looking at the correlation matrix of the predictors may reveal large pairwise correlations.
2. Looking at the variance inflation factors.

The variance inflation factors are calculated as follows: when an independent variable x_i, is regressed against all the other independent variables and the multiple coefficient of determination is R_i^2, the quantity $1/(1-R_i^2)$ is called the variance inflation factor for the parameter β_i (Mendenhall and Sincich, 2003). These variance inflation factors are

calculated for each numerical independent variable. Mendenhall and Sincich (2003) state that any value greater than 10 would mean that there is a collinearity problem.

3.3.4 Variable selection

Variable selection in generalized linear models is often done using a stepwise procedure or by best subset selection. Here the stepwise method is introduced for generalized linear models (Hosmer and Lemeshow, 2000).

Stepwise methods for generalized linear models

The forward stepwise variable selection method starts with no variables in the model and adds the most important variables sequentially. The backward stepwise variable selection method goes the other way around by starting with a model with all the variables and then sequentially deleting variables that provide little value in explaining the response. A stepwise procedure is based on a statistical algorithm that checks for the importance of variables. The steps are as follows:

Step 0 (select the best one variable model): Each possible variable is fitted individually and compared to the null model using a likelihood ratio test. The *p*-value for a significant variable must fall below a specific significance level. For example, with logistic regression, a significance level of between 0.15 and 0.20 is suggested. The variable with the smallest *p*-value below the significance level is chosen.

Step 1 (select the best two variable model): A generalized linear model is fitted containing the variable selected in step 0. Models are then fitted using the variable selected in step 0 and each of the other remaining models. These models are then compared to the model with the variable selected in step 0 using a likelihood ratio test. The variable with the smallest *p*-value is then chosen provided it is below the significance level.

Step 2: This procedure is continued until all variables are entered into the model or additional variables become insignificant.

Alternatively, the backward elimination procedure works by starting with all variables in the model, then removing the one that is least significant, then the next, etc until all the variables are significant.

This stepwise algorithm can also be conducted by comparing the AIC (Akaike information criterion) instead of using a likelihood ratio test.

These variable selection methods, however, become questionable with binary data. So it is better to consider variable selection using expert knowledge about which variables to include or not.

3.3.5 Logistic regression

Ntzoufras (2009) explains that data encountered with a binary response are often modelled with logistic regression. Logistic regression is a special case of the Generalized Linear Models (GLMs). For credit scoring data, a response $Y = 1$ represents a default or "bad" score and a response $Y = 0$ represents no default or "good" score. Logistic regression makes use of the canonical link function, $\ln(\frac{p}{1-p})$. The logistic regression model is given below

$Y_i \sim \text{binomial}(p_i, N_i)$, $\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = \boldsymbol{X}_{(i)}\boldsymbol{\beta}$ for $i = 1,2,\dots,n$. x_{ij} is the element in the *i*th row and *j*th column of the model matrix $\boldsymbol{X}$.

From this, the probability of default is given by

$$p_i = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}.$$

Other link parameters are also possible to model binary response data, for example the probit and clog-log links.

The likelihood for the logistic regression model is

$$L(y|\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{3.13}$$
$$= \prod_{i=1}^{n}\left(\frac{\exp\left(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}\right)}{1+\exp(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij})}\right)^{y_i}\left(1-\frac{\exp\left(\beta_0+\sum_{j=1}^{k}\beta_j x_{ij}\right)}{\exp(1+\beta_0+\sum_{j=1}^{k}\beta_j x_{ij})}\right)^{1-y_i}.$$

Estimation of the parameters for logistic regression can be done using the IRWLS procedure.

Parameter interpretation

The parameters in logistic regression have an interpretation in terms of odds and odds ratios. Odds is defined as the relative probability of success ($Y = 1$) compared to the probability of failure ($Y = 0$) when the data is binomial (Ntzoufras, 2009). Thus,

$$odds = \frac{p}{1-p}$$

and the logistic regression model can be rewritten as

$$Y_i \sim \text{binomial}\left(\frac{odds_i}{1+odds_i}, N_i\right), \ln(odds_i) = \beta_0 + \sum_{j=1}^{k}\beta_j x_{ij} = \boldsymbol{X}_{(i)}\boldsymbol{\beta}.$$

Odds provides a number to multiply the probability of failure by in order to calculate the probability of success. β_j can be interpreted as follows: a unit increase in x_{ij} with all the other x_{ij}'s held fixed increases the log-odds of success by β_j or increases the odds of success by e^{β_j}. This interpretation is a major advantage of logistic regression as no such simple interpretation exists for other link functions such as the probit.

In credit scoring, a success corresponds to a default or bad applicant. Thus, the log-odds of success is the log-odds of default in the context of credit scoring. Therefore, β_j can be interpreted as follows: a unit increase in x_{ij} with all the other x_{ij}'s held fixed increases the log-odds of default by β_j or increases the odds of default by e^{β_j}. A positive value for β_j thus increases the odds of default as x_{ij} increases, while a negative value for β_j decreases the odds of default as x_{ij} increases.

Assessment of fit

Dobson and Barnett (2008) state that one way of assessing the fit of a model is to compare it with a model with the maximum number of parameters. The model with the maximum number of parameters is called the saturated model and has the same number of parameters as covariate patterns (i.e. observations with the same values of all the variables). The saturated model tells us no more than the actual data and is often non-informative (Faraway, 2006). However, we can use the saturated model to compare prospective models. The difference between the log-likelihood for the full model and model under consideration gives the likelihood ratio statistic, known as the deviance

$$D = 2[l(\widehat{\boldsymbol{\beta}}_{max}) - l(\widehat{\boldsymbol{\beta}})].$$

The deviance for the binomial model is now derived. This follows from Dobson and Barnett (2008). From Equation (3.13) the likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \exp[y_i \ln(p_i) - y_i \ln(1-p_i) + n_i \ln(1-p_i) + \ln \binom{n_i}{y_i}].$$

This in term means the log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} [y_i \ln(p_i) - y_i \ln(1-p_i) + n_i \ln(1-p_i) + \ln \binom{n_i}{y_i}]. \quad (3.14)$$

From this we find the maximum likelihood estimate for p_i. Now, differentiating and equating to zero we have

$$\frac{\partial l}{\partial p_i} = \frac{y_i}{p_i} + \frac{y_i}{1-p_i} - \frac{n_i}{1-p_i} = 0$$
$$\Rightarrow \frac{y_i}{p_i} + \frac{y_i}{1-p_i} = \frac{n_i}{1-p_i}$$
$$\Rightarrow \frac{(1-p_i)y_i + p_i y_i}{p_i(1-p_i)} = \frac{n_i}{1-p_i}$$
$$\Rightarrow \frac{y_i}{p_i} = n_i$$

which leads to the maximum likelihood estimate

$\hat{p}_i = \frac{y_i}{n_i}$.

Now, the maximum value of the log-likelihood function Equation (3.14) is

$l(\widehat{\boldsymbol{\beta}}_{max}) = \sum_{i=1}^{n}[y_i \ln\left(\frac{y_i}{n_i}\right) - y_i \ln\left(\frac{n_i - y_i}{n_i}\right) + n_i \ln\left(\frac{n_i - y_i}{n_i}\right) + \ln\binom{n_i}{y_i}]$.

For any other model with number of parameters less than the number of covariate patterns, let $\hat{y}_i = n_i\hat{p}_i$ denote the fitted values. Then, the log-likelihood evaluated at these values is

$l(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}[y_i \ln\left(\frac{\hat{y}_i}{n_i}\right) - y_i \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right) + n_i \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right) + \ln\binom{n_i}{y_i}]$.

Therefore, the deviance for the Binomial model is

$$
\begin{aligned}
D &= 2\left[l(\widehat{\boldsymbol{\beta}}_{max}) - l(\widehat{\boldsymbol{\beta}})\right] \\
&= 2\sum_{i=1}^{n}[[y_i \ln\left(\frac{y_i}{n_i}\right) - y_i \ln\left(\frac{n_i - y_i}{n_i}\right) + n_i \ln\left(\frac{n_i - y_i}{n_i}\right) + \ln\binom{n_i}{y_i}] - \sum_{i=1}^{n}[y_i \ln\left(\frac{\hat{y}_i}{n_i}\right) \\
&\qquad - y_i \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right) + n_i \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right) + \ln\binom{n_i}{y_i}]] \\
&= 2\sum_{i=1}^{n}[\ y_i \ln\left(\frac{y_i}{n_i}\right) - y_i \ln\left(\frac{\hat{y}_i}{n_i}\right) - y_i \ln\left(\frac{n_i - y_i}{n_i}\right) + y_i \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right) + n_i \ln\left(\frac{n_i - y_i}{n_i}\right) \\
&\qquad - n_i \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right)] \\
&= 2\sum_{i=1}^{n}[\ y_i \left(\ln\left(\frac{y_i}{n_i}\right) - \ln\left(\frac{\hat{y}_i}{n_i}\right)\right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{n_i}\right) - (n_i - y_i) \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right)] \\
&= 2\sum_{i=1}^{n}[\ y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)[\ln\left(\frac{n_i - y_i}{n_i}\right) - \ln\left(\frac{n_i - \hat{y}_i}{n_i}\right)]] \\
&= 2\sum_{i=1}^{n}[\ y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\left[\ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]. \qquad (3.15)
\end{aligned}
$$

This deviance has a chi-squared distribution with degrees of freedom equal to the number of covariate patterns less the number of parameters. The deviance can, therefore, be used in a hypothesis test to assess the fit of a model. However, when the outcome is binary, i.e. when y_i takes on the values zero or one, this goodness-of-fit measure is no longer useful.

There is also a Hosmer-Lemeshow statistic which tries to overcome the problem of a goodness-of-fit statistic for binary data (Hosmer and Lemeshow, 2000). However, its use is still questionable.

Classification

Logistic regression models a binary outcome. The objective is often to classify.

In order to perform classification, Hosmer and Lemeshow (2000) explain that a cut-off point, c, must be defined. The estimated probabilities from the logistic regression model are compared to this cut-off point. If the estimated probability exceeds c, we let the derived variable be equal to 1; if the estimated probability is less than c, we let the derived variable be equal to 0.

Classification tables, according to Hosmer and Lemeshow (2000) are a good way to summarize the results of a fitted logistic regression model. The outcome variable y is cross classified with a dichotomous variable whose values are derived from the estimated logistic probabilities.

The predictive performance of the logistic regression model is probably the best way to assess the fit of the model. The way to do this will be to split a data set into a training and a test set. The model will be estimated on the training set and its performance will be tested on a test set with a certain cut-off probability, c. A classification table will then be established and the error rate of the model can be used as a measure of how well the model fits (Table 3.1).

Table 3.1 Classification table of the predictive performance of the logistic regression model.

		Predicted	
		Good	Bad
Actual	Good	w	x
	Bad	y	z

From Table 3.1, a number of facts can be established.

- $w + x + y + z$ represents the number of applicants in the test set.
- $x + z$ is the number of applicants classified as bad. This is the number of applicants who were rejected in their application for credit.
- $w + y$ is the number of applicants classified as good. This is the number of applicants who were accepted in their application for credit.
- w is the number of applicants correctly classified as good and z is the number of applicants correctly classified as bad.
- x is the number of applicants classified as bad but are in fact good. This number represents missed out profits for the financial institution.
- y is the number of applicants classified as good but are in fact bad. This number represents bad debts and losses in income for the financial institution.
- The total error probability of the classification is $(y + x)/(w + x + y + z)$. This value must be small. It also gives an indication of the goodness-of-fit of the model.
- For the applicants that will be accepted by the financial institution, the error probability is $y/(w + y)$. This is the error rate realized by the bank. Thus, it is very important that y is as small as possible.
- A cut-off probability c, needs to be found which minimizes the classification error.

The choice of the cut-off probability c, is often a subjective choice. For lower c, more applicants will be classified as bad. For higher c, the more applicants will be classified as good. A lower cut-off probability means that the financial institution is more risk averse as opposed to one with a higher cut-off probability. Because the error rate realized by the financial institution is greatly affected by how large y is, it is important that the error among the bad applicants is minimized as well as the total error.

The optimal cut-off probability can be found by using a validation set. The classification error can be determined for different cut-off probabilities. The cut-off probability which gives the lowest classification error on the validation set will be chosen and used in further analysis.

3.3.6 Bayesian logistic regression

Bayesian inference for the logistic regression model requires priors on the model parameters. Wilhelmsen *et al*. (2009) and Ziemba (2005) both use normally distributed priors for the model parameters and represented as follows

$$\pi(\beta_i) = N(\beta_{0i}, \sigma_i^2). \tag{3.16}$$

The posterior distribution is proportional to the product of the prior distribution and likelihood, $\pi(\beta|y) \propto L(y|\beta)\pi(\beta)$. Therefore, from Equations (3.13) and (3.16), we have

$$\pi(\beta|y) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\beta_j-\beta_{0j})^2}{2\sigma_j^2}\right) \tag{3.17}$$

or

$$\pi(\beta|y) = \frac{\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{(\beta_j-\beta_{0j})^2}{2\sigma_j^2})}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{(\beta_j-\beta_{0j})^2}{2\sigma_j^2}) d\beta} \tag{3.18}$$

when the normalising constant is included.

The form of this posterior distribution, Equation (3.17), suggests that the prior does not belong to a conjugate family. There is in fact no conjugate prior for the Bayesian logistic regression model. The normalising constant, the integral in the denominator (Equation (3.18)) cannot be calculated explicitly. In this situation simulation methods need to be used in order to obtain the posterior distributions of the parameters. Markov Chain Monte Carlo (MCMC) methods are used where a Markov chain is generated with a stationary distribution equal to the posterior distribution of the vector **β**.

3.4 Monte Carlo Methods

Simulation has greatly improved on the scope of Bayesian inference. Markov Chain Monte Carlo (MCMC) methods allow for sampling from a non-standard distribution. Therefore, Bayesian inference can be done in a wide range of posterior distribution forms, for example Equation (3.18). The idea is to generate a Markov chain whose limiting (stationary) distribution is equal to the posterior distribution. This section will describe simulation techniques, provide an introduction to Markov chains and then explain the role and purpose of Markov Chain Monte Carlo.

3.4.1 Monte Carlo simulation

In Bayesian inference, simulation is needed to evaluate integrals. In order to do this, it is essential that random data can be generated. The generation of random variables and all other Monte Carlo methods are reliant on the generation of uniform random variables on the interval (0,1).

Uniform random number generation

There are many methods to produce pseudo uniform random numbers as shown in Kroese *et al*. (2011). These generators include Linear congruential, Multiple-recursive, Matrix congruential, Modulo 2 linear etc. The function for the multiple-recursive generator is as follows:

$x_{n+1} = ax_n \mod m;$

$u_{n+1} = \frac{x_{n+1}}{m}$

where a and m are positive integers and $ax_n \mod m$ means that ax_n is divided by m and the remainder is taken as the next value x_{n+1}. To use the generator, only a starting number

is thus needed. This starting number is called the seed. Once the desired number of random numbers have been generated, each number is divided by m. This results in uniform random numbers on the interval $(0,1)$.

According to Kroese *et al*. (2011) two excellent generators that have very good performance are:

- Combined multiple-recursive generators.
- Twisted general feedback shift register generators.

Luckily, these very good generators are what are used in computer programs and statistical software. For example the program MATLAB uses the twisted general feedback shift register generator.

Random variable generation

Two common methods for random variable generation are the inverse transform method and accept-reject algorithm.

Inverse-transform method

Kroese *et al*. (2011) introduces the inverse-transform method as follows:

Let X be a random variable with cumulative distribution function (cdf) $F(x) = P(X \leq x)$. Since F is a non-decreasing function, the inverse function can be defined as

$$F^{-1}(y) = \begin{cases} \min\{x: F(x) \geq y\} & \text{if } y > 0 \\ -\infty & \text{if } y = 0 \end{cases}.$$

Now if we have a random variable U from a uniform distribution on $(0,1)$, i.e $U \sim unif(0,1)$, then the cdf of the inverse transform $F^{-1}(U)$ is given by

$$P(F^{-1}(U) \leq x) = P\big(U \leq F(x)\big) = F(x).$$

Thus, in order to generate a random variable X with cumulative distribution function $F(x)$, we generate U from $unif(0,1)$ and then make the transformation $x = F^{-1}(u)$. Therefore, we have the following algorithm

- 1. Generate U from $unif(0,1)$;
- 2. Return $X = F^{-1}(U)$.

This is used for sampling random variables from continuous distributions. Obviously, this method only works when we can determine and evaluate the inverse of the cdf F.

Accept-reject method

The inverse transform method is of no use when one cannot obtain the inverse of the cumulative distribution function. A more general method is the accept-reject method which can be used to sample from more general distributions.

According to Greenberg (2008), the accept-reject method can be used to simulate random variables from a density function $f(x)$ when it is possible to simulate values from another density $g(x)$, and if a number $M \geq 1$ can be found such that $f(x) \leq Mg(x)$ for all x. The density $g(x)$ is called the instrumental or candidate density. In order to simulate random variables X from $f(x)$ Robert and Casella (2010) state that first, we independently generate $Y \sim g(x)$ and $U \sim unif(0,1)$. Then, if

$$U \leq \frac{1}{M}\frac{f(Y)}{g(Y)},$$

we set $X = Y$. If not we discard Y. This leads to the accept-reject algorithm

- 1. Generate Y from $g(x)$;
- 2. Generate U from $unif(0,1)$, independently of Y;
- 3. Accept $X = Y$ if $U \leq \frac{1}{M}\frac{f(Y)}{g(Y)}$, else reject Y;
- 4. Return to 1.

Following Robert and Casella (2010), the cdf of the accepted random variable $P\left(Y \leq x \middle| U \leq \frac{f(Y)}{Mg(Y)}\right)$ is exactly the cdf of X. That is,

$$P\left(Y \le x \middle| U \le \frac{f(Y)}{Mg(Y)}\right) = \frac{P\left(Y \le x, U \le \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \le \frac{f(Y)}{Mg(Y)}\right)}$$

$$= \frac{\int_{-\infty}^{x}[\int_{0}^{\frac{f(y)}{Mg(y)}} du]g(y)dy}{\int_{-\infty}^{\infty}[\int_{0}^{\frac{f(y)}{Mg(y)}} du]g(y)dy}$$

$$= \frac{\int_{-\infty}^{x} \frac{f(y)}{Mg(y)} g(y)dy}{\int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y)dy}$$

$$P\left(Y \le x \middle| U \le \frac{f(Y)}{Mg(Y)}\right) = \frac{\int_{-\infty}^{x} f(y)dy}{\int_{-\infty}^{\infty} f(y)dy} = \int_{-\infty}^{x} f(y)dy = P(Y \le y).$$

The output is, therefore, exactly distributed from $f(x)$.

Considering the efficiency of this method, we note that the probability of accepting a point is given by.

$$P(\text{accept}) = P\left(U \le \frac{f(y)}{Mg(y)}\right) = \int_{-\infty}^{\infty} (\int_{0}^{\frac{f(y)}{Mg(y)}} 1du)\, g(y)dy$$

$$= \int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y)dy$$

$$= \int_{-\infty}^{\infty} \frac{1}{M} f(y)dy = \frac{1}{M}.$$

This implies that we should choose an M as small as possible in order to maximize the probability of acceptance. The algorithm is efficient when g is as close to f as possible. Maximizing the probability of acceptance is important because as Greenberg (2008, p 67) states, "rejected values use computer time without adding to the sample", therefore, deceasing efficiency.

Monte Carlo integration

Monte Carlo integration is a statistical technique for approximating integrals. It uses simulation to obtain an estimate of the integral which has a mean and a variance. One method of Monte Carlo Integration is the sample mean approach. This method is described below for the estimation of the integral, $I = \int_a^b f(x)dx$. The following approach is discussed in Suess and Trumbo (2010).

Now, if $X \sim unif(a,b)$ then $E\big(f(X)\big) = \int_a^b \left(\frac{1}{b-a}\right) f(x)dx = \frac{1}{b-a}\int_a^b f(x)\,dx$. Therefore,

$$\int_a^b f(x)dx = (b-a)E\big(f(x)\big).$$

The integral $\int_a^b f(x)dx$ can, therefore, be approximated by

$$\theta = \frac{b-a}{N}\sum_{k=1}^{N} f(u_k) \tag{3.19}$$

where $u_1, u_2, \dots, u_N$ are random numbers from $unif(a,b)$. The mean and variance of this estimator is derived as follows:

$$E(\theta) = E(\frac{b-a}{N}\sum_{k=1}^{N} f(u_k)) = \frac{b-a}{N}NE(f(U)) = \frac{b-a}{N}N\frac{1}{b-a}\int_a^b f(u)du = I.$$

Therefore, the estimator in Equation (3.19) is an unbiased estimator for the integral, $I = \int_a^b f(x)\,dx$. Now, for the variance

$$\begin{aligned} var(\theta) &= var(\frac{b-a}{N}\sum_{k=1}^{N} f(u_k)) \\ &= \frac{(b-a)^2}{N^2}Nvar\big(f(U)\big) \\ &= \frac{(b-a)^2}{N}[(E(f(U)^2) - \Big(E\big(f(U)\big)\Big)^2] \\ &= \frac{(b-a)^2}{N}[\frac{1}{b-a}\int_a^b \big(f(u)\big)^2 du - (\frac{1}{b-a}\int_a^b f(u)du)^2] \\ &= \tfrac{1}{N}(b-a)[\int_a^b \big(f(u)\big)^2 dx - \Big(\int_a^b f(u)dx\Big)^2]. \end{aligned} \tag{3.20}$$

So, we have that $var(\theta) \propto \frac{1}{N}$.

Importance sampling

Importance sampling is used to reduce the variance of a Monte Carlo estimate of an integral. From Equation (3.20) the standard deviation of an estimator for the integral, *I* is $std(\theta) \propto \frac{1}{\sqrt{N}}$. Thus, the standard deviation of the estimator decreases as N increases, but at a decreasing rate. This means that if we increase the number of random points from $N = 10^2$ to $N = 10^4$ points, the standard deviation is improved from the order of $\frac{1}{10}$ to $\frac{1}{100}$. Therefore, quite a large number of random points are needed to obtain a noticeable improvement in accuracy. Importance Sampling aims to improve the standard deviation of a Monte-Carlo estimate. The idea is as follows as seen in Robert and Casella (2004).

Consider a density $p(x)$ on $[a, b]$ with the property that $p(x) > 0$ whenever $f(x) \neq 0$. Then

$\int_a^b f(x)dx = \int_a^b \frac{f(x)}{p(x)} p(x)dx = E_p(\frac{f(X)}{p(X)})$ if $X \sim p(x)$.

Therefore, in order to obtain an estimate for $\int_a^b f(x)\,dx$ using importance sampling, we sample $x_1, x_2, \ldots, x_N$ from $p(x)$and estimate

$$\int_a^b f(x)dx \approx \frac{1}{N}\sum_{k=1}^{N} \frac{f(x_k)}{p(x_k)}. \tag{3.21}$$

The new estimator is given by $\vartheta = \frac{1}{N}\sum_{k=1}^{N} \frac{f(X_k)}{p(X_k)}$ and the variance of ϑ is given by

$$\begin{aligned} var(\vartheta) &= var\left(\frac{1}{N}\sum_{k=1}^{N}\frac{f(X_k)}{p(X_k)}\right) \\ &= \frac{1}{N^2} N var\left(\frac{f(X)}{p(X)}\right) \\ &= \frac{1}{N}\left[E\left(\frac{f(X)}{p(X)}\right)^2 - \left(E\left(\frac{f(X)}{p(X)}\right)\right)^2\right] \\ &= \frac{1}{N}\left[\int_a^b \frac{\left(f(x)\right)^2}{\left(p(x)\right)^2} p(x)dx - \left(\int_a^b f(x)dx\right)^2\right]. \end{aligned}$$

To minimize the variance, we need to minimize the first term $E\left(\frac{f(X)}{p(X)}\right)^2$. Using Jensen's inequality

$$E\left(\frac{f(X)}{p(X)}\right)^2 \geq [E\left(\frac{|f(X)|}{p(X)}\right)]^2 = (\int_a^b \frac{|f(x)|}{p(x)} p(x)dx)^2 = (\int_a^b |f(x)|dx)^2 \quad (3.22)$$

which is a lower bound and does not depend on the choice of $p(x)$.

Theorem 3.1 If $\vartheta = \frac{1}{N}\sum_{k=1}^{N} \frac{f(X_k)}{p(X_k)}$ where $X_1, X_2, \ldots, X_N$ are i.i.d. from a density $p(x)$ such that $p(x) > 0$ whenever $f(x) > 0$ and $\int_a^b p(x)dx = 1$ then

1. $E(\vartheta) = \int_a^b f(x)dx.$
2. $var(\vartheta)$ is minimized if $p(x) = \frac{|f(x)|}{\int_a^b |f(x)|dx}$.

Proof:

1. $E(\vartheta) = E\left(\frac{1}{N}\sum_{k=1}^{N} \frac{f(X_k)}{p(X_k)}\right) = E\left(\frac{f(X)}{p(X)}\right) = \int_a^b \frac{f(x)}{p(x)} p(x)dx = \int_a^b f(x)dx.$
2. From Equation (3.22) $(\int_a^b |f(x)|dx)^2$ is the lower bound for $E\left(\frac{f(X)}{p(X)}\right)^2$ which is what we want to minimize. Now, if $p(x) = \frac{|f(x)|}{\int_a^b f(x)dx}$,

$$\begin{aligned} E\left(\frac{f(X)}{p(X)}\right)^2 &= \int_a^b (f(x))^2 \frac{\int_a^b |f(x)|dx}{|f(x)|} dx \\ &= \int_a^b |f(x)|dx \int_a^b |f(x)|dx \\ &= (\int_a^b |f(x)|dx)^2. \end{aligned}$$

Thus, for this choice of $p(x)$ we are at the lower bound and variance is minimized. Now, the practical use of Theorem 3.1 is very limited. This is because we need to know the integral $\int_a^b |f(x)|dx$ which is for $f(x) \geq 0$ the same as $\int_a^b f(x)dx$. But $\int_a^b f(x)dx$ is what we are looking to estimate in the first place. This theorem does, however, help us to choose a good $p(x)$. We should try to achieve $\frac{f(x)}{p(x)} \approx$ constant. Therefore, we should sample more

points in regions where $f(x)$ is large. Thus, the "important" parts of the integral will be estimated better. This is why the method is called importance sampling.

3.4.2 Markov chains

An overview of Markov chains is given in this section. The simulation methods described previously cannot easily be applied in all cases. Monte Carlo integration and importance sampling can be applied when we are dealing with standard distributions. However, when we face a non-standard distribution (such as the case with Bayesian logistic regression) the previous simulation techniques cannot easily be used to obtain samples from any posterior distribution. If they are used, they are subject to major practical difficulties. Markov Chain Monte Carlo (MCMC) methods provide a way out.

Markov Chain Monte Carlo methods have greatly improved the scope for Bayesian inference (Robert and Casella, 2004; Greenberg, 2008). Because MCMC relies on Markov chains, they are now introduced with both discrete and continuous state spaces.

Discrete state space

The definition of a Markov chain is as follows:

Definition 3.1 Let $(X_0, X_1, X_2, \dots)$ be a stochastic process indexed by t (often time) that takes values in the finite set $S = \{1, 2, \dots, s\}$ (finite state space) or $S = \{1, 2, \dots\}$ (infinite state space). If the Markov property

$$P(X_{t+1} = j | X_t = k, X_{t-1} = k_{t-1}, \dots, X_1 = k_1, X_0 = k_0) = P(X_{t+1} = j | X_t = k) = p_{kj}$$

holds true for all states j, k, k_{t-1}, … , k_1, $k_0 \in S$ and all time steps $t = 0, 1, 2, \dots$, then $(X_0, X_1, X_2, \dots)$ is called a Markov chain.

Therefore, the current state of a Markov chain only affects the next state. The p_{kj} are transition probabilities. These transition probabilities do not depend on the time t. Since the p_{kj} are probabilities, we have $p_{kj} \geq 0$ and since the process remains in S

$$\sum_{j=1}^{s} p_{kj} = 1.$$

The transition probabilities are very important and it is useful to collect these transition probabilities in a matrix. The $s \times s$ transition probability matrix is given by

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & & p_{2s} \\ \vdots & & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{pmatrix}.$$

The *i*th row of P, specifies the distribution of the process at $t + 1$, given that it is in state i at t. For example, p_{22} represents the probability of going to state 2 given that it is in state 2.

We now consider multi-step transition probabilities $p_{kj}^{(n)}$, which are defined as follows

$$p_{kj}^{(n)} = P(X_n = j | X_0 = k) = P(X_{m+n} = j | X_m = k).$$

The calculation of multi-step transition probabilities is made easy from the following Chapman-Kolmogorov lemma:

Lemma 3.1 Let $(X_0, X_1, X_2, \dots)$ be a Markov chain with state space $S = \{1,2,3, \dots\}$. Then, we have for the multi-step transition probabilities

$$p_{kj}^{(m+n)} = \sum_{i \in S} p_{ki}^{(m)} p_{ij}^{(n)}.$$

Proof:

$$\begin{aligned} p_{kj}^{(m+n)} &= P(X_{m+n} = j | X_0 = k) \\ &= \sum_{i \in S} P(X_{m+n} = j, X_m = i | X_0 = k) \end{aligned}$$

$$= \sum_{i \in S} \frac{P(X_{m+n} = j, X_m = i, X_0 = k)}{P(X_0 = k)} \frac{P(X_m = i, X_0 = k)}{P(X_m = i, X_0 = k)}$$

$$= \sum_{i \in S} P(X_{m+n} = j | X_m = i, X_0 = k) P(X_m = i | X_0 = k).$$

Now, using the Markov property we obtain

$$p_{kj}^{(m+n)} = \sum_{i \in S} P(X_{m+n} = j | X_m = i) P(X_m = i | X_0 = k).$$

This implies,

$$p_{kj}^{(m+n)} = \sum_{i \in S} p_{ki}^{(m)} p_{ij}^{(n)}.$$

The Chapman-Kolmogorov lemma can also be written in matrix form

$$\boldsymbol{P}^{m+n} = \boldsymbol{P}^m \boldsymbol{P}^n.$$

We now turn to a discussion of the classification of states.

Some states will be visited over and over again, while others will only be visited a finite number of times and never visited again. Let, for a state $i \in S$

$$T_i = \min\{n \geq 1 : X_n = i\}.$$

Thus, T_i is the time of first visit to state i. Also let

$$f_i = P(T_i < \infty | X_0 = i)$$

which is the probability that the Markov chain will return to state i once it started there. There are two possible cases for the f_i's:

1. $f_i = 1$. This means that we are certain we will continuously return to state i (over and over again). Such a state is called recurrent and will be visited infinitely many times.
2. $f_i < 1$. This means that there is a positive probability of never returning to state i. Such a state is called transient which will only be visited a finite amount of times.

We say that a state j is accessible from state i if there is $n \geq 0$ such that $p_{ij}^{(n)} > 0$ and write $i \to j$. State j is accessible from state i if with a finite number of steps we can come

from state i to j. Also, if $i \to j$ and $j \to i$ we then say that the states i and j communicate and expressed as $i \leftrightarrow j$. It can be shown that the communication relation is an equivalence relation between the states of S. This means, we have for all states $i, j, k \in S$

- $i \leftrightarrow i$ (reflexivity);
- If $i \leftrightarrow j$ then also $j \leftrightarrow i$ (symmetry);
- If $i \leftrightarrow j$ and $j \leftrightarrow k$, then also $i \leftrightarrow k$ (transitivity).

A Markov chain is known as irreducible if there is only one communication class. This means that all states communicate (the process can reach any other state with positive probability). This would imply that for an irreducible Markov chain with finite state space, that all the states are recurrent.

The distribution $\pi = (\pi_1, \pi_2, \dots)$ is called a stationary (or invariant or limiting) distribution if $\pi = \pi P$. This limiting distribution, $\pi = \lim_{t\to\infty} \pi^t$ exists if the Markov chain is irreducible and all states are aperiodic (the greatest common divisor of the sets $A_i = \left\{t \geq 1 : p_{ii}^{(t)} > 0\right\}$ is one).

We now introduce Markov chains for a continuous state space.

Continuous state space

It is now assumed we have a stochastic process $(X_0, X_1, X_2, \dots)$ with discrete time but a continuous state space $S \subseteq \mathbb{R}^d$ and that all distributions have densities.

For Markov chains with continuous state space the transition probabilities $P(X_{t+1} = x_{t+1} | X_t = x_t)$ are always zero. Therefore, looking at specific points $x \in S$, when defining transition probabilities, is not helpful. Thus, subsets $A \subseteq S$ are considered. This leads to the following definition:

Definition 3.2 Let $(X_0, X_1, X_2, \dots)$ be a stochastic process with continuous state space $S \subseteq \mathbb{R}^d$. If for all $A \subseteq S$ and all states $x_0, x_1, \dots, x_t \in S$ we have

$$P(X_{t+1} \in A | X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_1 = x_1, X_0 = x_0) = P(X_{t+1} \in A | X_t = x_t),$$

then we call the stochastic process a Markov chain with continuous state space (Robert and Casella, 2004).

We will always assume that we can determine the transition probabilities

$P(X_{t+1} \in A | X_t = x_t)$ using a transition kernel $K: S \times S \to \mathbb{R}_{\geq 0}$ by

$$P(X_{t+1} \in A | X_t = x_t) = \int_{x_{t+1} \in \mathrm{A}} K(x_t, x_{t+1}) dx_{t+1}.$$

A transition kernel has the following properties

- $K(x_t, x_{t+1}) \geq 0$ for all $x_t, x_{t+1} \in S$;
- $\int_{x_{t+1} \in \mathrm{S}} K(x_t, x_{t+1}) dx_{t+1} = 1.$

It can be shown that the two-step transition probability is given by

$$P(X_2 \in A | X_0 = x_0) = P(X_2 \in A, X_1 \in S | X_0 = x_0)$$
$$= \int_{x_2 \in A} \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) dx_1 dx_2.$$

Therefore, the two-step transition kernel is

$K^{(2)}(x_0, x_2) = \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) dx_1.$

We can generalize this to T-step transitions (multi-step transitions).

$P(X_T \in A | X_0 = x_0) =$

$\int_{x_T \in A} \int_{x_{T-1} \in S} \dots \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) \dots K(x_{T-1}, x_T)\, dx_1 \dots dx_{T-1} dx_T.$

Hence, we have the T-step transition kernel

$K^{(T)}(x_0, x_T) = \int_{x_{T-1} \in S} \dots \int_{x_1 \in S} K(x_0, x_1) K(x_1, x_2) \dots K(x_{T-1}, x_T)\, dx_1 \dots dx_{T-1}.$

Therefore, we have

$P(X_T \in A | X_0 = x_0) = \int_{x_T \in A} K^{(T)}(x_0, x_T)\, dx_T.$

The Chapman-Kolmogorov lemma is then also true with a countable state space

$$K^{(T+S)}(x_0, x_{T+S}) = \int_{x_T \in S} K^{(T)}(x_0, x_T) K^{(S)}(x_T, x_{T+S}) dx_T.$$

The concept of an irreducible Markov chain (as discussed under discrete state spaces) is the same for the continuous state space. Thus, the definitions of recurrent and transient Markov chains, communication and aperiodic also apply to the continuous case.

The concept of a stationary distribution for a Markov chain with continuous state space is now discussed.

We assume that for a Markov chain $(X_0, X_1, X_2, \dots)$ with transition kernel K, the distribution of X_t has the density p_t. Now, if $p_{t+1}(x_{t+1})$ is the density of X_{t+1} then

$p_{t+1}(x_{t+1}) = p(x_{t+1}|x_t) p_t(x_t) = p_t(x_t) K(x_t, x_{t+1}).$

Therefore, for the distribution of X_{t+1} for any $A \subseteq S$

$P(X_{t+1} \in A) = \int_{x_{t+1} \in A} p_{t+1}(x_{t+1}) dx_{t+1} = \int_{x_{t+1} \in A} \int_{x_t \in S} p_t(x_t) K(x_t, x_{t+1}) dx_t d\, x_{t+1}.$

A density for X_{t+1} is therefore given by

$p_{t+1}(x_{t+1}) = \int_{x_t \in S} p_t(x_t) K(x_t, x_{t+1}) dx_t.$

We then have the following definition:

Definition 3.3 A probability distribution π with density p is called a stationary distribution for a Markov chain $(X_0, X_1, X_2, \dots)$ with transition kernel K if

$$p(y) = \int_{x \in S} p(x) K(x, y) dx$$

for all $y \in S$ except on a set $A \subseteq S$ with $\pi(A) = 0$.

Such a distribution is called an invariant distribution. We now move onto the so-called detailed balance condition.

Lemma 3.3 Let $(X_0, X_1, X_2, \dots)$ be a Markov chain with transition kernel K. If, for a density function p, we have the detailed balance condition

$p(y)K(y,x) = p(x)K(x,y)$ for all $x, y \in S$

then p is the density of a stationary distribution of the Markov chain.

Proof:

We have

$$\int_{x\in S} p(x)K(x,y)dx = \int_{x\in S} p(y)K(y,x)\,dx = p(y)\int_{x\in S} K(y,x)dx = p(y).$$

Definition 3.4 Let $(X_0, X_1, X_2, \dots)$ be a Markov chain with continuous state space S. Let v be a probability distribution on S. The Markov chain is called v-irreducible if for all $x_0 \in S$ and all $A \subseteq S$ with $v(A) > 0$ there is $T \in \mathbb{N}$ such that

$P(X_T \in A | X_0 = x_0) = \int_{y\in A} K^{(T)}(x_0, y)dy > 0.$

If $T = 1$ then the Markov chain is called strongly v-irreducible. This property of a Markov chain implies that any set with a positive probability $v(A) > 0$ can be visited from any $x_0 \in S$ in finite time. Thus, if this property holds, all states communicate.

Now, let $\eta_A = \sum_{t=1}^{\infty} 1_A(X_t)$ denote the number of visits of the Markov chain in the set A.

Definition 3.5 Let $(X_0, X_1, X_2, \dots)$ be a Markov chain and let $A \subseteq S$. We then call

- the set A recurrent if for all $x_0 \in A$ we have $E(\eta_A | X_0 = x_0) = \infty$.

- the Markov chain recurrent if it is v-irreducible for some probability distribution v and whenever $v(A) > 0$, then A is recurrent.

A stronger definition of recurrence is now given.

Definition 3.6 Let $(X_0, X_1, X_2, \dots)$ be a Markov chain and let $A \subseteq S$. We then call

- the set A Harris-recurrent if for all $x_0 \in A$ we have $P(\eta_A = \infty | X_0 = x_0) = 1$.
- the Markov chain Harris-recurrent if it is v–irreducible for some probability distribution v and whenever $v(A) > 0$, then A is Harris-recurrent.

Lemma 3.4 Let $(X_0, X_1, X_2, \dots)$ be a Markov chain with stationary distribution π (with density p). If $X \sim p$ and if the Markov chain is π-irreducible and recurrent, then for any integrable function $h: S \to \mathbb{R}$ we have (with probability 1)

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X_t) = \int_S h(x)p(x)dx = E_p(h(X))$$

for almost all starting values $X_0 = x_0$. If the Markov chain is Harris-recurrent, then the equation holds for all $x_0 \in S$.

3.4.3 Markov chain Monte Carlo

Markov chain Monte Carlo constructs a Markov chain that has as its stationary distribution, the target distribution. It does this by constructing an irreducible Markov chain, which ensures that most of the Markov chains resulting from an MCMC algorithm are recurrent or even Harris-recurrent. As explained, Harris recurrence ensures that the Markov chain converges to its stationary distribution for every starting value instead of almost every starting value. Thus, we need Harris recurrence to ensure that the MCMC algorithm converges. MCMC algorithms construct a transition kernel which results in a

Markov chain which is recurrent and converges to the target distribution. A general principle to do this is the Metropolis-Hastings (MH) algorithm. The Gibbs sampler is a special case of the MH algorithm.

Metropolis-Hastings algorithm

We wish to construct a Markov chain which has its stationary distribution equal to the target distribution.

The results from the previous section are now used to construct a transition kernel, $K(x, y)$, that has an invariant density equal to the target density. We consider this in the continuous case. The Metropolis-Hastings algorithm is a general algorithm for sampling from any form of posterior distribution.

The Metropolis-Hastings (MH) algorithm has two ingredients: Lemmas 3.3 and 3.4. Lemma 3.4 essentially means that we can sample dependent samples from a Markov chain and we can use $\frac{1}{T}\sum_{t=1}^{T} h(X_t)$ to estimate $E_p(h(X))$.

We now use Lemma 3.3 (detailed balance condition). A transition kernel that holds true for this lemma is known as a reversible kernel and results in a stationary distribution. This lemma can help in finding a kernel that has the desired target distribution. Following Chib and Greenberg (1995), we make an irreversible kernel reversible. If a kernel is not reversible for some pair (x, y) we may have

$$p(x)K(x, y) > p(y)K(y, x). \tag{3.23}$$

We wish to make this inequality an equality. In this case there are more moves from x to y than y to x. In order to achieve an equality, before we make a move from x to y we impose a probability $\alpha(x, y) < 1$ with which such a move will be accepted. This probability $\alpha(x, y)$ must be such that $\alpha(x, y)p(x)K(x, y) = p(y)K(y, x)$. This means that

$$\alpha(x, y) = \min\left(\frac{p(y)K(y, x)}{p(x)K(x, y)}, 1\right).$$

This ensures that the detailed balance condition holds. If, on the other hand, we have

$$p(x)K(x,y) < p(y)K(y,x)$$

then we multiply the right-hand-side by $\alpha(y,x) < 1$ and obtain

$$\alpha(y,x) = \min\left(\frac{p(x)K(x,y)}{p(y)K(y,x)}, 1\right).$$

This then leads to the Metropolis-Hastings algorithm:

- 1. Choose a transition kernel q with $q(x,y) > 0$ for all states $x, y \in S$;
- 2. Start at $t = 0$ with some arbitrary state $X_t = x_t \, \epsilon \, S$;
- 3. If $X_t = x_t$, generate a random variable $Y \sim q(x_t, \cdot)$ and $U \sim unif(0,1)$;
- 4. If $Y = y$ (and $X_t = x_t$) set

  $$X_{t+1} = \begin{cases} y \; if \; U \leq \alpha(x_t, y) \\ x_t \qquad\qquad else \end{cases};$$
- 5. $t = t + 1$ and return to 3.

Here $\alpha(x,y) = \min(\frac{p(y)q(y,x)}{p(x)q(x,y)}, 1)$.

This Metropolis-Hastings algorithm is the principle algorithm which is used with Bayesian logistic regression.

The transition kernel, q is the proposal kernel. There is considerable freedom in choosing the proposal kernel. However, care still needs to be taken in order to choose particularly useful ones. For example, when the proposal kernel does not "explore" the whole state space of $p(x)$ then certain values will not be sampled. There are two common choices for the proposal kernel which lead to the independence sampler and the random walk sampler.

The choice of the proposal kernel affects the acceptance rates of the algorithm. According to Ntzoufras (2009), the variance of the proposal controls the convergence speed of the algorithm. Small variances of the proposal kernel will result in high acceptance rates, but low convergence since the algorithm will need a large number of iterations to explore the entire parameter space. Conversely, a high variance will result in low acceptance rates and a highly correlated sample. The optimal acceptance rate is between 20% and 40% (Ntzoufras, 2009). For models with a large number of parameters the acceptance rate should be towards the lower bound, for a univariate model the acceptance rate should be towards the upper bound. The way to obtain the acceptance rate in this range is by tuning

the variance of the proposal kernel. Metropolis-Hastings algorithms include a tuning parameter. This parameter is "tuned" so that the acceptance rate is between 20-40%.

Independence sampler

If the proposal $q(x,y)$ does not depend on y, that is $q(x,y) = g(y)$ for all x then the acceptance probability is

$$\alpha(x,y) = \min\left(\frac{p(y)g(x)}{p(x)g(y)}, 1\right).$$

The independence sampler is very similar to the accept-reject method in Section 3.4.1. Like the accept-reject method, it is important that the proposal kernel, g, is close to the target f to allow for efficient simulation. However, the independence sampler produces dependent samples. Also, if there is a constant M such that $p(x) \leq Mg(x)$, then the expected acceptance rate is at least $1/M$ when the Markov chain is stationary. The proof is as follows:

$$E\big(\alpha(X_t,Y)\big) = E\left(\min\left(\frac{p(Y)g(X_t)}{p(X_t)g(Y)}, 1\right)\right) = \int\int \min\left(\frac{p(y)g(x)}{p(x)g(y)}, 1\right) p(x)g(y)dxdy$$

$$= \int\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}>1} p(x)g(y)dxdy + \int\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\leq 1} \frac{p(y)g(x)}{p(x)g(y)} p(x)g(y)dxdy$$

$$= \int\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}>1} p(x)g(y)dxdy + \int\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\leq 1} p(y)g(x)dxdy$$

$$= 2\int\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\geq 1} p(x)g(y)dxdy$$

$$\geq \frac{2}{M}\int\int_{(x,y):\frac{p(y)g(x)}{p(x)g(y)}\geq 1} p(x)g(y)dxdy$$

$$= \frac{2}{M}\int\int_{(x,y):\frac{p(y)}{g(y)}\geq\frac{p(x)}{g(x)}} p(x)p(y)dxdy\,.$$

If $h(x) = \frac{p(x)}{g(x)}$, the last integral is $P(h(U) \geq h(V))$ for U, V independent with distribution $p(x)$. Therefore, it is as likely that $h(U) \geq h(V)$ as that $h(V) \geq h(U)$ if U, V are independent and identically distributed. Thus,

$$E\big(\alpha(x,y)\big) \geq \frac{2}{M}\frac{1}{2} = \frac{1}{M}.$$

Random walk sampler

The other common choice is to use the current simulated value to generate the next value. In that way, the neighbourhood of the current value of the Markov chain is explored. A proposal kernel which allows this is the symmetrical kernel $q(x,y) = q(y,x)$. This leads to acceptance probability

$$\alpha(x,y) = \min\left(\frac{p(y)}{p(x)}, 1\right).$$

A proposed value $Y = y$ is then accepted with probability one if $p(y) \geq p(x)$. Thus, points y that are more likely (according to p) than the previous x_t will always be accepted. However, we also accept points which are less likely with a certain probability. Thus, making use of the previous x_t explores S in a more local way.

Markov chain Monte Carlo diagnostics

The Markov chain, from which we take samples, needs to have converged to the target distribution. If the Markov chain has not converged, samples will be taken which are not from the desired target distribution. In order to make sure that samples are taken only from the stationary distribution, a burn-in period is used (Ntzoufras, 2009). The burn-in period is the number of samples which are eliminated to ensure we only sample from the stationary distribution.

Dobson and Barnett (2008) state that a way to assess the convergence of a Markov chain is by looking at time series (trace) plots. These graphs plot the history of the Markov chain. A chain that has converged should be stable and show a reasonable degree of randomness between iterations.

Another way to assess convergence is by looking at the autocorrelation function (ACF) of the chain. Ideally we would want samples to be independent, but with MCMC algorithms this cannot happen. We therefore accept some autocorrelation. If the ACF values are low it indicates that the Markov chain has converged successfully.

Geweke (1992) proposed a diagnostic test for assessing the convergence of the mean of each parameter. He considers the simulated Markov chain (obtained from the MCMC output) as a time series and applies a z-test to check whether the means from two different subsamples are equal. These subsamples come from the beginning and end of the generated chain. Typically, the first 10% of the chain is used as the beginning sample and the last 50% is used as the end sample. Using this z-test, parameters with $|z| > 2$ indicate evidence of significant differences between the means of the first and last set of iterations and means non-convergence of the chain.

Chapter 4: Results

4.1 Initial Data Analysis

Analysis is now illustrated on a real life home equity data set. This data set was first analysed by Wielenga *et al*. (1999). Here, the methodology described in Section 3.1 is performed.

The data set contains the loan performance of 5,960 home equity loans. The target (dependent) variable is a dummy variable indicating whether or not a default occurred during the duration of the loan. If a default occurred the value is one; if no default occurred the value is zero. The data set consists of 12 input (independent) variables. The variables are summarized in Table 4.1.

Table 4.1 Variable type and description for each variable in the data set.

Variable	Model Role	Variables Type	Description
BAD	Target	Categorical - Nominal	1 = defaulted on loan 0 = paid back loan
REASON	Input	Categorical - Nominal	HomeImp = home improvement DebtCon = debt consolidation
JOB	Input	Categorical - Nominal	Six occupational categories
LOAN	Input	Numerical - Continuous	Amount of loan request
MORTDUE	Input	Numerical - Continuous	Amount due on existing mortgage
VALUE	Input	Numerical - Continuous	Value of current property
DEBTINC	Input	Numerical - Continuous	Debt-to-income ratio
YOJ	Input	Numerical - Continuous	Years at present job
DEROG	Input	Numerical - Discrete	Number of major derogatory reports
CLNO	Input	Numerical - Discrete	Number of trade lines
DELINQ	Input	Numerical - Discrete	Number of delinquent trade lines
CLAGE	Input	Numerical - Continuous	Age of oldest trade line in months
NINQ	Input	Numerical - Discrete	Number of recent credit inquiries

The target variable is a binary variable consisting of ones and zeros. There are 1,189 ones and 4,771 zeros in the target variable. This means that close to 20% of the applicants defaulted during the duration of the loan or became seriously delinquent.

For the input variables, there are two categorical variables. The other ten input variables are all numerical of which four are discrete and six are continuous. Summary statistics for the input numerical variables are given in Table 4.2.

Table 4.2 Summary statistics for the numerical input variables.

Variable	Minimum	Median	Mean	Maximum	SD	NAs
LOAN	1100.00	16300.00	18608.00	89900.00	11207.48	0
MORTDUE	2063.00	65019.00	73761.00	399550.00	45095.37	518
VALUE	8000.00	89236.00	101776.00	855909.00	54728.24	112
DEBTINC	0.5245	34.8183	33.7799	203.3121	7.9514	1267
YOJ	0.000	7.000	8.922	41.000	7.596	515
DEROG	0.0000	0.0000	0.2546	10.0000	0.5795	708
CLNO	0.00	20.00	21.30	71.00	9.39	222
DELINQ	0.0000	0.0000	0.4494	15.0000	0.8096	580
CLAGE	0.0	173.5	179.8	1168.2	82.8	308
NINQ	0.000	1.000	1.186	17.000	1.549777	510

NAs represent the number of missing values for each variable. There are many missing values in all of the variables except for LOAN. In particular, most of the missing values occur in the DEBTINC variable with 1,267 missing values. If all the missing values in the data are ignored, the proportion of applicants who defaulted goes down to 8.9%. Because there is such a large number of missing values and the proportion of bad applicants decreases when they are ignored, an estimation technique was used to estimate the missing values.

Bar plots for the categorical variables are given in Figure 4.1.

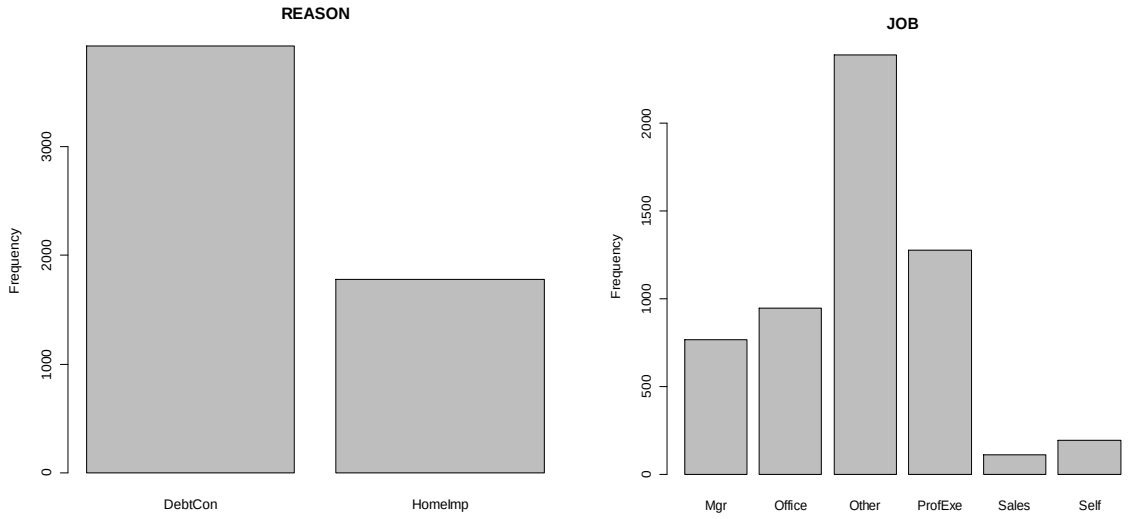


Fig. 4.1 Bar plots of REASON and JOB.

Histograms and box plots for the numerical variables are given in Figures 4.2 to 4.11 (Histogram on the left-hand-side and box plot on the right-hand-side).

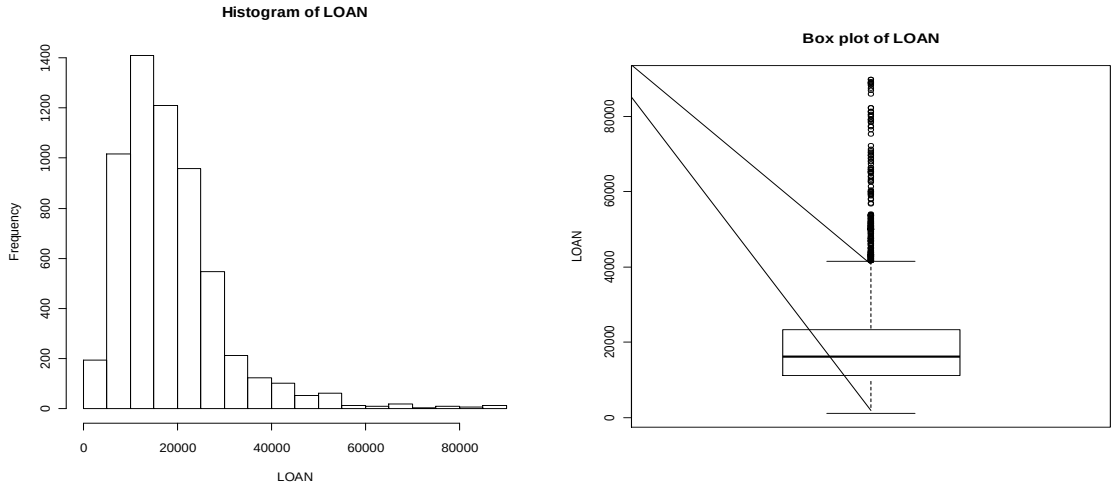


Fig. 4.2 Histogram and Box plot of LOAN.

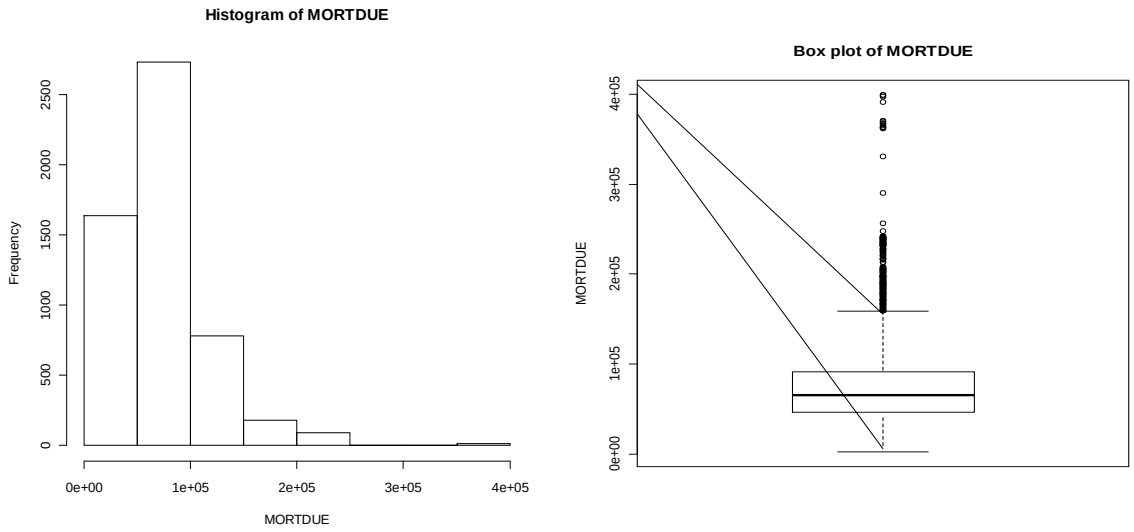


Fig. 4.3 Histogram and Box plot of MORTDUE.

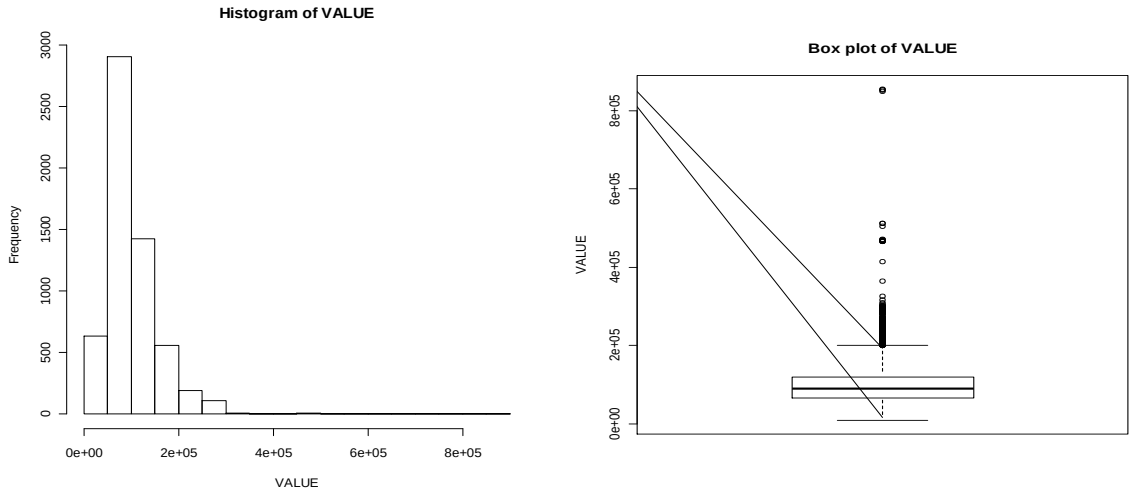


Fig. 4.4 Histogram and Box plot of VALUE.

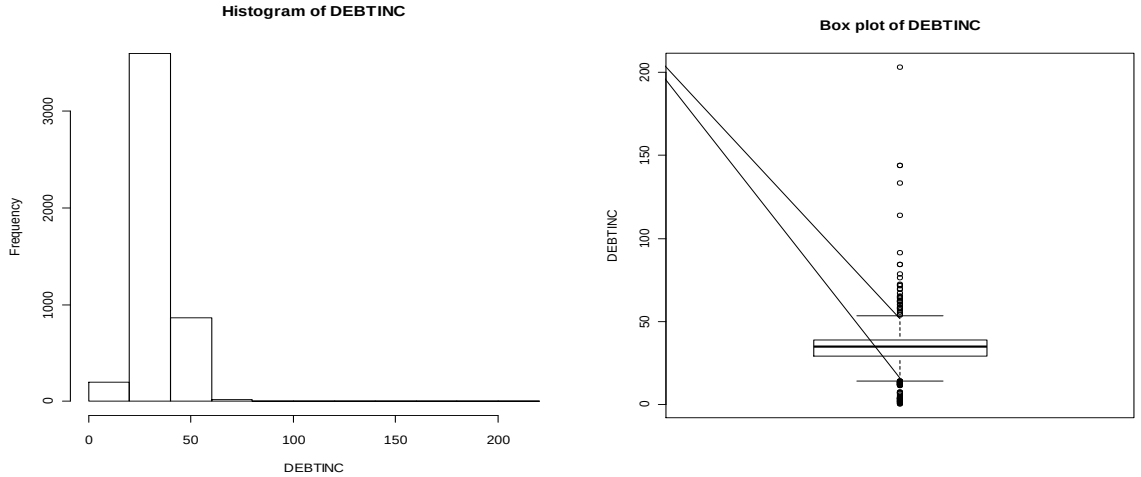


Fig. 4.5 Histogram and Box plot of DEBTINC.

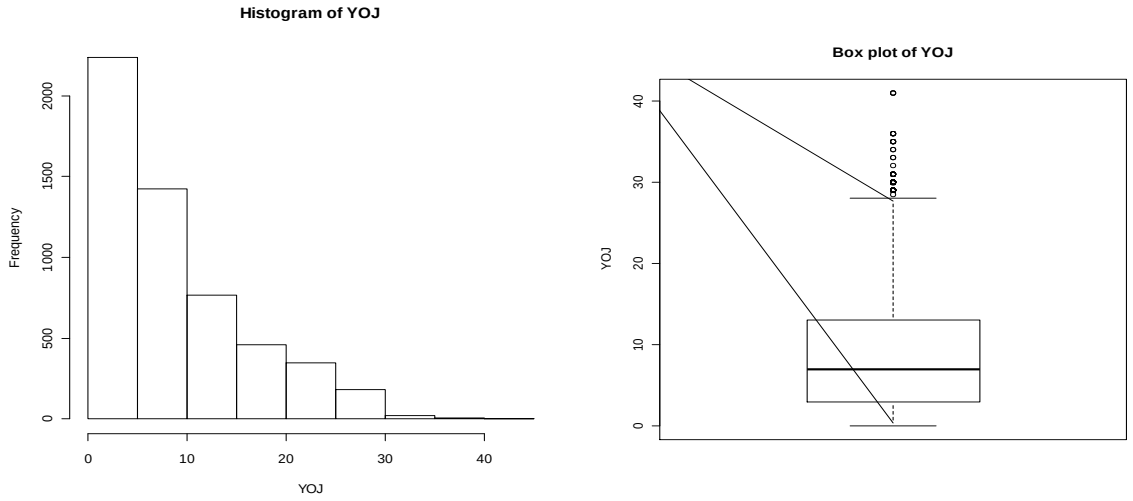


Fig. 4.6 Histogram and Box plot of YOJ.

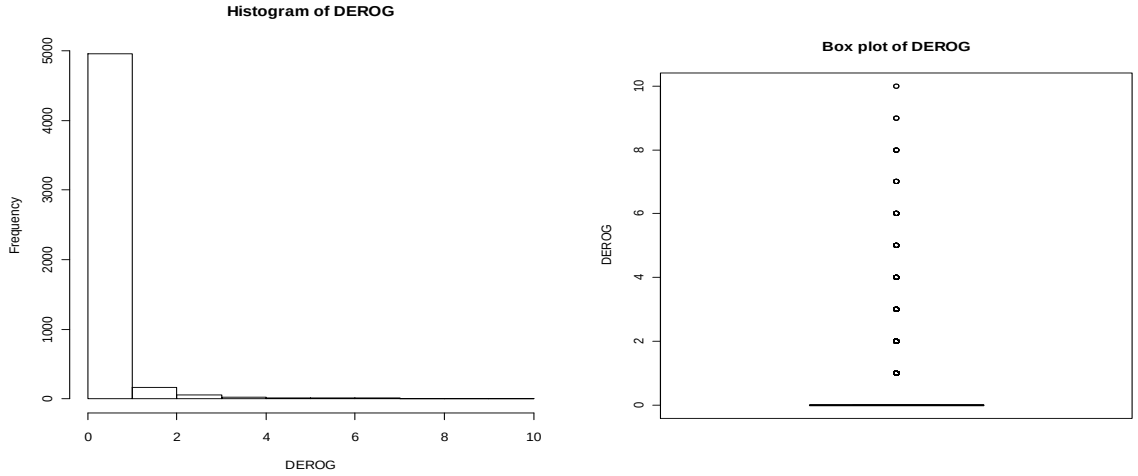


Fig. 4.7 Histogram and Box plot of DEROG.

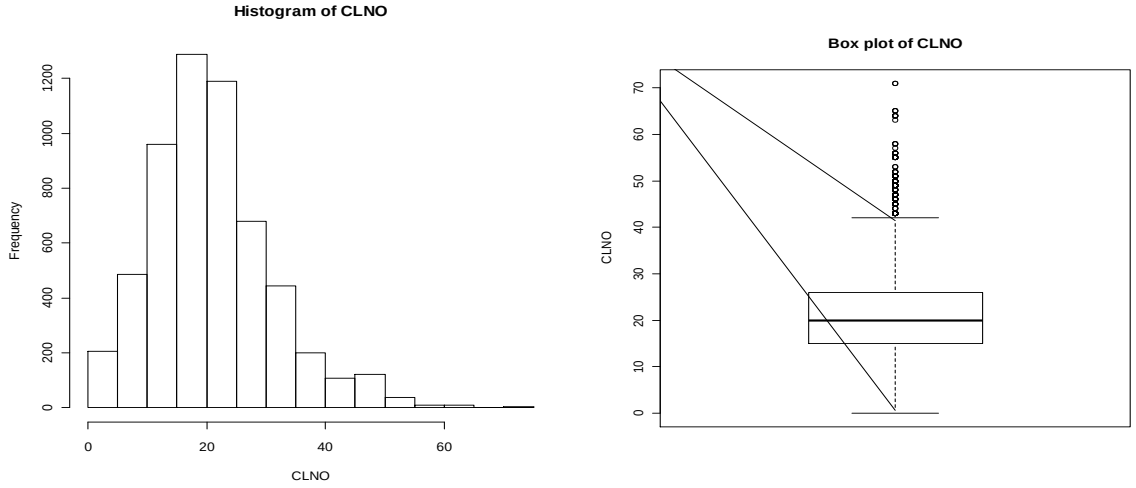


Fig. 4.8 Histogram and Box plot of CLNO.

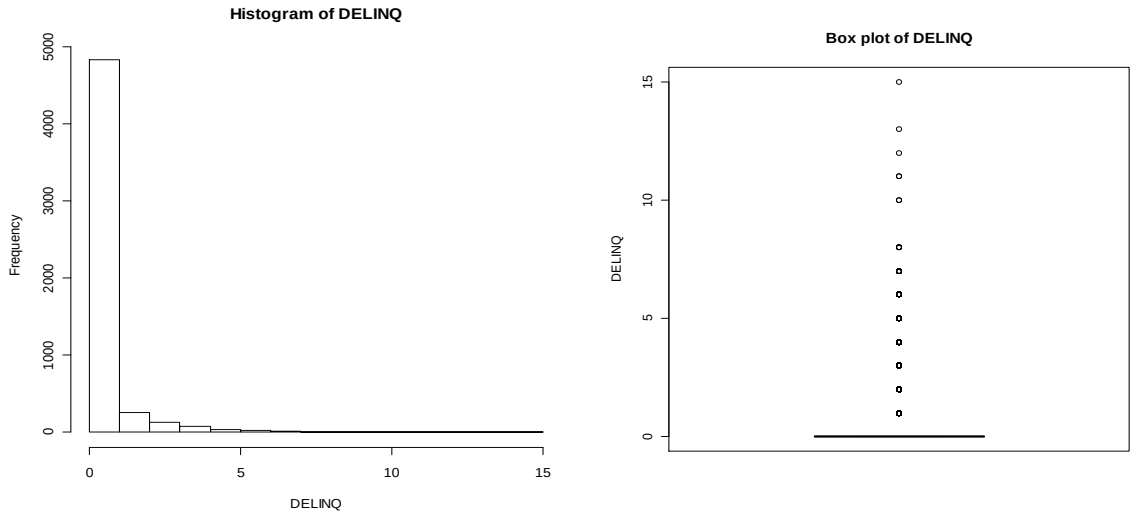


Fig. 4.9 Histogram and Box plot of DELINQ.

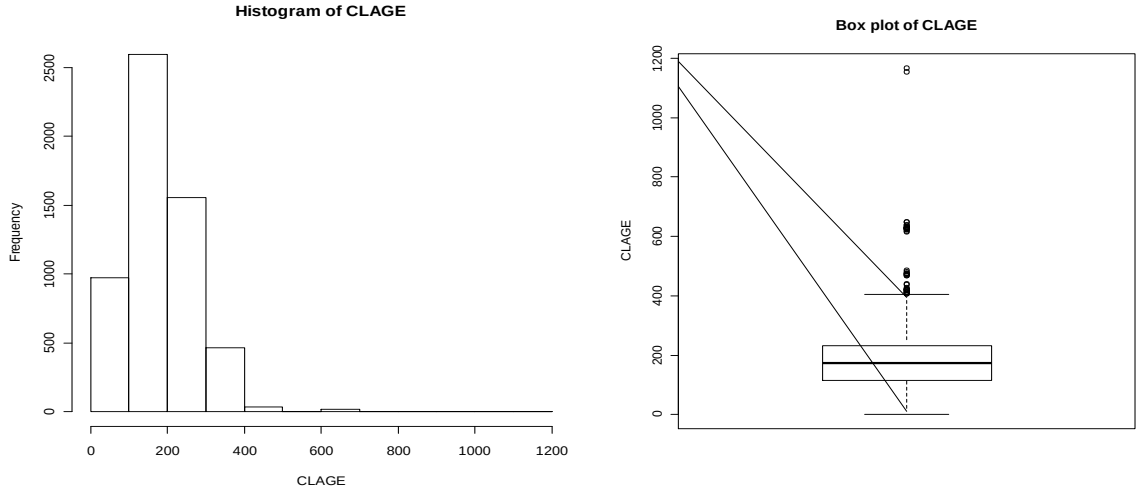


Fig. 4.10 Histogram and Box plot of CLAGE.

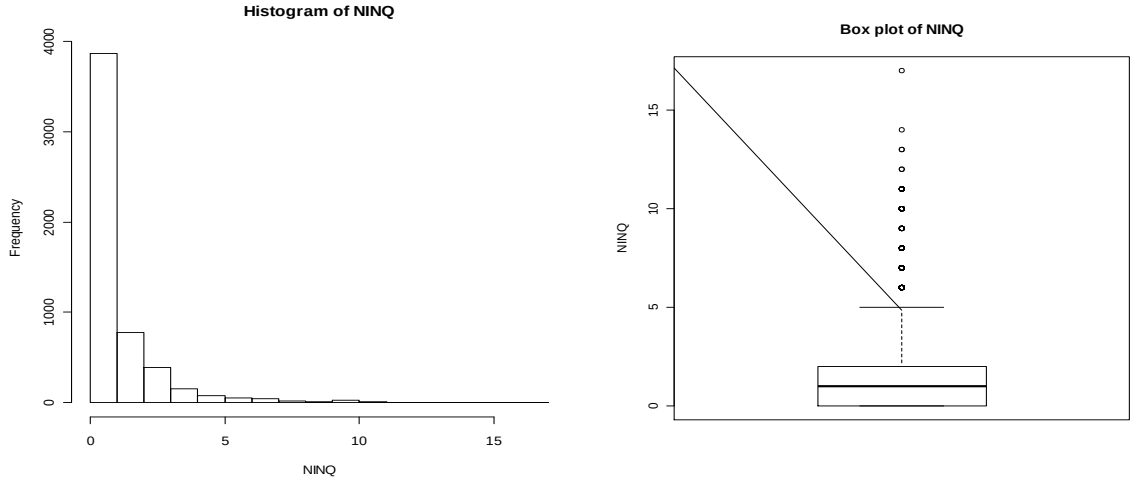


Fig. 4.11 Histogram and Box plot of NINQ.

From Figures 4.2 to 4.11, in the majority of cases there appears to be a number of outliers towards the right-tails. This might result in the variables being more positively skewed than they should be. For example, the variable MORTDUE appears to have a number of outliers in the right-tail. For the variables DELINQ and DEROG, the majority of the values are zero. The question now arises whether these are legitimate outliers or whether they are outliers caused by errors in recording. This is addressed when the models are fitted.

The data set was randomly split into four sets:

- The "old" data set contains 2,759 observations of which 565 are bad.
- The "validation" data set contains 549 observations of which 109 are bad.
- The "new" data set contains 566 observations of which 114 are bad.
- The "test" data set contains 1,662 observations of which 340 are bad.

The missing values in the data set were replaced by the mean for each variable when the target variable (BAD) was equal to 1 and when it was equal to 0. The missing values were thus replaced by two means for each variable.

4.2 Logistic Regression Model on "old" Data

A logistic regression model was fitted on the "old" data. This model is the model fitted on the available data in the home country. Six Fisher scoring iterations were needed for the algorithm, used to fit the model, to converge. The estimated parameters of the model are given in Table 4.3.

Table 4.3 Logistic regression model fitted on the "old" data.

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-7.19E+00	5.64E-01	-12.765	< 2e-16	Significant
LOAN	-2.37E-05	6.50E-06	-3.642	0.000271	Significant
MORTDUE	-3.71E-06	2.28E-06	-1.625	0.104238	Insignificant
VALUE	3.03E-06	1.60E-06	1.902	0.057212	Insignificant
REASONHomeImp	2.03E-01	1.35E-01	1.504	0.132632	Insignificant
JOBOffice	-6.82E-01	2.25E-01	-3.038	0.002382	Significant
JOBOther	1.72E-02	1.79E-01	0.096	0.923139	Insignificant
JOBProfExe	4.76E-02	2.10E-01	0.227	0.820586	Insignificant
JOBSales	4.02E-01	4.25E-01	0.948	0.343111	Insignificant
JOBSelf	4.02E-01	3.80E-01	1.057	0.290496	Insignificant
YOJ	-1.62E-02	9.14E-03	-1.768	0.077093	Insignificant
DEROG	7.34E-01	8.06E-02	9.098	< 2e-16	Significant
DELINQ	8.04E-01	6.42E-02	12.53	< 2e-16	Significant
CLAGE	-5.22E-03	8.65E-04	-6.038	1.56E-09	Significant
NINQ	1.37E-01	3.20E-02	4.272	1.94E-05	Significant
CLNO	-2.82E-02	6.79E-03	-4.148	3.36E-05	Significant
DEBTINC	1.91E-01	1.38E-02	13.868	< 2e-16	Significant

There are a number of significant variables at the 5% level of significance. This indicates that many of the variables included in the model are significant in explaining whether an applicant will be good or bad. The residual deviance of the model is 1,866.7 with 2,742 degrees of freedom.

Interpretation is now given for the parameters of LOAN, DEROG and DEBTINC.

- The parameter of LOAN is -2.37E-05 and is significant at the 5% significance level. LOAN represents the amount of loan request. A unit increase in LOAN with all other variables held fixed, means that there will be a 2.37E-05 decrease in the log-odds of default.
- The parameter of DEROG is 7.34E-01 and is significant at the 5% significance level. DEROG represents the number of major derogatory reports. A unit increase in DEROG

with all other variables held fixed, means that there will be a 7.34E-01 increase in the log-odds of default.

- The parameter of DEBTINC is 1.91E-01 and is significant at the 5% significance level. DEBTINC represents the debt to income ratio of the applicant. A unit increase in DEBTINC with all other variables held fixed, means that there will be a 1.91E-01 increase in the log-odds of default.

In order to check the adequacy of the model, collinearity of the independent variables, outliers and influential observations are considered. The correlation matrix of the numerical independent variables is given in Table 4.4.

From this correlation matrix, we see that there are no large pair-wise correlations. The largest correlation is 0.78 between VALUE and MORTDUE. Worrying correlations will occur with the correlation between two variables is greater than 0.9. The variance inflation factors for each numerical variable are given in Table 4.5.

Table 4.4 Correlation matrix of numerical independent variables on the "old" data.

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
LOAN	1.00	0.22	0.33	0.08	0.00	-0.06	0.09	0.04	0.05	0.02
MORTDUE	0.22	1.00	0.78	-0.09	-0.04	-0.01	0.14	0.00	0.31	0.09
VALUE	0.33	0.78	1.00	-0.01	-0.02	-0.01	0.16	-0.02	0.27	0.08
YOJ	0.08	-0.09	-0.01	1.00	-0.05	0.04	0.22	-0.06	0.03	-0.05
DEROG	0.00	-0.04	-0.02	-0.05	1.00	0.16	-0.09	0.18	0.04	0.06
DELINQ	-0.06	-0.01	-0.01	0.04	0.16	1.00	0.03	0.06	0.14	0.12
CLAGE	0.09	0.14	0.16	0.22	-0.09	0.03	1.00	-0.11	0.27	-0.05
NINQ	0.04	0.00	-0.02	-0.06	0.18	0.06	-0.11	1.00	0.07	0.14
CLNO	0.05	0.31	0.27	0.03	0.04	0.14	0.27	0.07	1.00	0.13
DEBTINC	0.02	0.09	0.08	-0.05	0.06	0.12	-0.05	0.14	0.13	1.00

Table 4.5 Variance inflation factors (VIF) of numerical independent variables on the "old" data.

Variable	VIF
LOAN	1.151227
MORTDUE	2.720736
VALUE	2.817692
YOJ	1.089154
DEROG	1.0712
DELINQ	1.066063
CLAGE	1.174603
NINQ	1.071983
CLNO	1.226312
DEBTINC	1.059467

From Table 4.5, we see there are no large variance inflation factors, which indicates that there is no serious problem with collinearity in the "old" data.

Outliers and influential observations in the model are now considered. The following plots are considered for the presence of outliers and influential observations: half-normal plots of the residuals, leverages and Cook's distance statistics. The half-normal plot of the residuals is given in Figure 4.12.

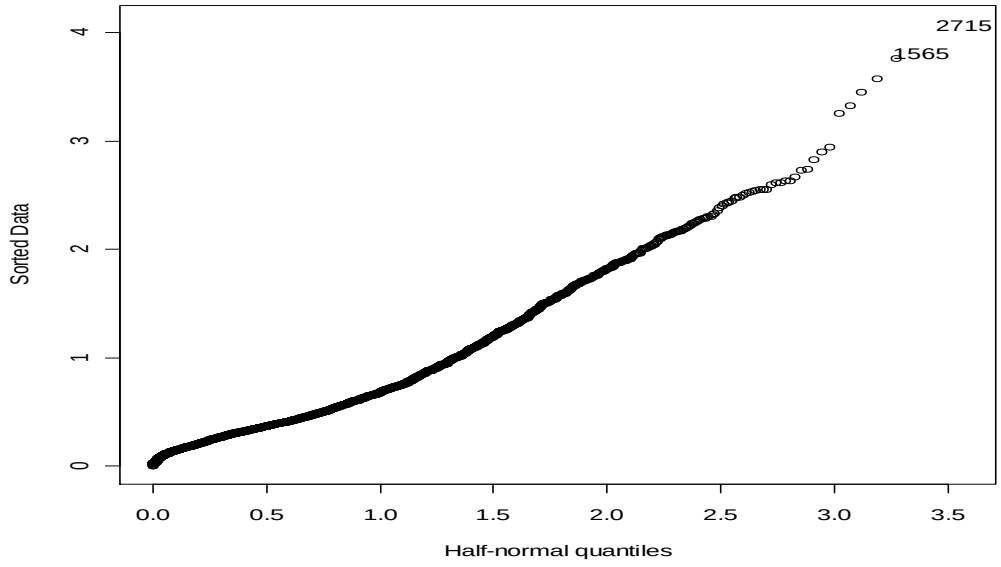


Fig. 4.12 Half-normal plot of residuals for the "old" data.

From Figure 4.12, there does not appear to be any sign of outliers. The half-normal plot of the leverages is given in Figure 4.13.

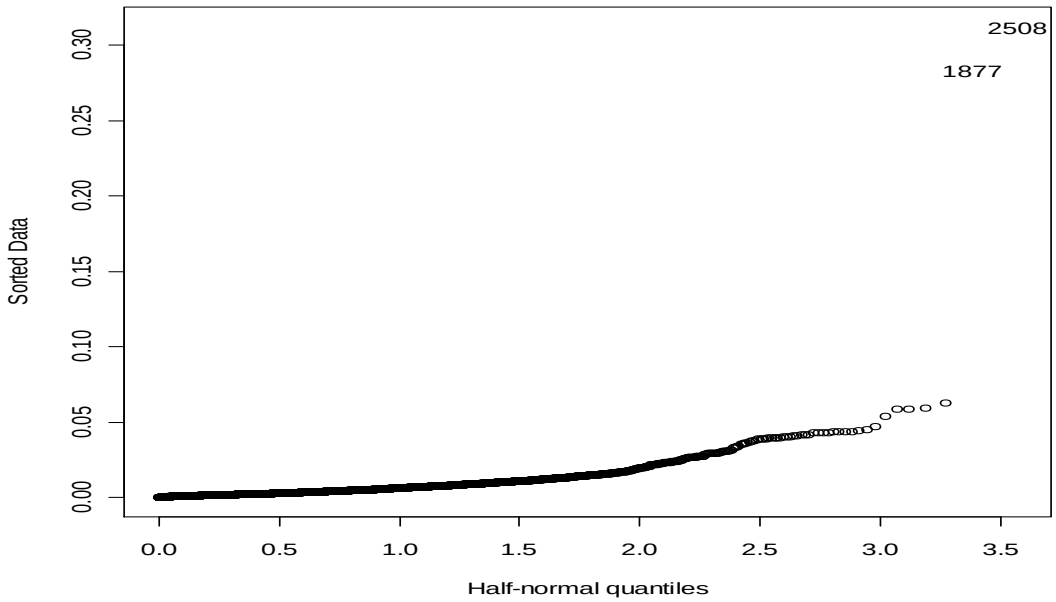


Fig. 4.13 Half-normal plot of leverages for the "old" data.

Figure 4.13 indicates that observations numbered 1877 and 2508 may have the potential to affect the fit of the model.

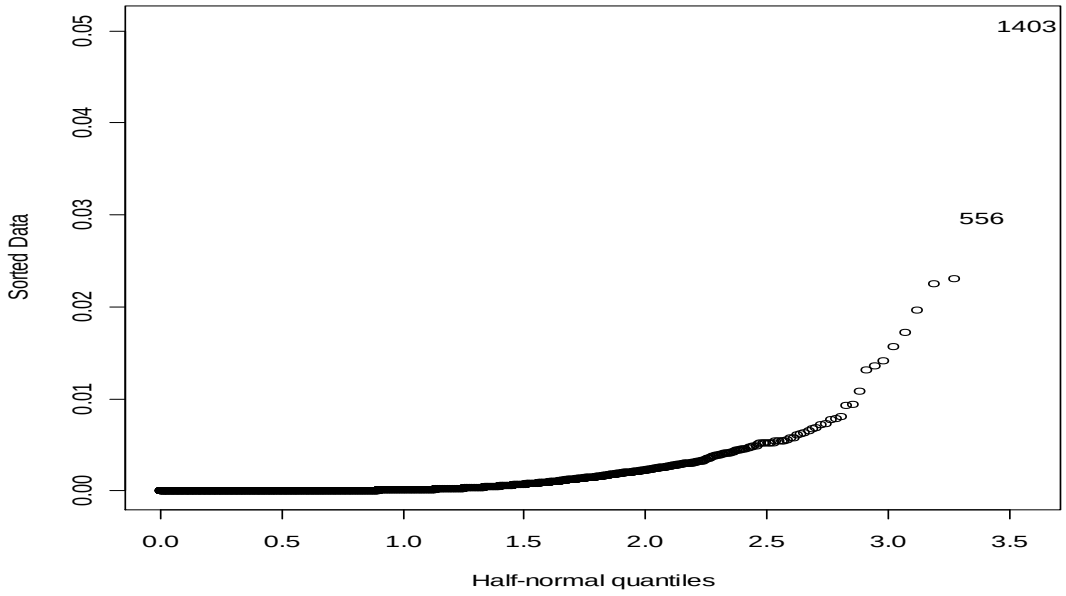


Figure 4.14 Half-normal plot of the Cook's distance statistics for the "old" data.

The half-normal plot of the Cook's distance statistics is given in Figure 4.14. This plot indicates that observations numbered 556 and 1403 may be influential. A logistic regression

model was then fitted on the "old" data excluding observations numbered 556, 1403, 1877 and 2508. The estimated parameters of this model were then compared to the model with no removed observations.

Table 4.6 Comparison of model coefficients when possible leverage and influential observations are either included or excluded from the "old" data.

Variable	Coefficients including	Coefficients excluding
(Intercept)	-7.19E+00	-7.09E+00
LOAN	-2.37E-05	-2.21E-05
MORTDUE	-3.71E-06	-2.94E-06
VALUE	3.03E-06	1.49E-06
REASONHomeImp	2.03E-01	2.07E-01
JOBOffice	-6.82E-01	-6.61E-01
JOBOther	1.72E-02	6.31E-03
JOBProfExe	4.76E-02	9.34E-02
JOBSales	4.02E-01	4.42E-01
JOBSelf	4.02E-01	4.68E-01
YOJ	-1.62E-02	-1.42E-02
DEROG	7.33E-01	7.33E-01
DELINQ	8.04E-01	8.09E-01
CLAGE	-5.22E-03	-5.95E-03
NINQ	1.37E-01	1.36E-01
CLNO	-2.81E-02	-2.56E-02
DEBTINC	1.91E-01	1.91E-01

Table 4.6 shows the difference in the parameters when possible leverage and influential observations are removed. The first column gives the model parameters with all observations in the "old" data and the second column gives the model parameters when the possible leverage and influential observations are removed. Looking at Table 4.6, the differences in the estimated parameters between the models are minimal. Therefore, the possible leverage and influential points will not be removed from the "old" data.

4.3 Determining an Optimal Cut-off Probability

We wish to minimize the error rate of the classification performance of the model. Because the data is skewed towards the good loans, we need to make sure that the model classifies the bad loans sufficiently. In order to do this, the minimization of the following function is proposed

$$function_{error} = \alpha\,(error_{total}) + (1 - \alpha)(error_{\text{bad}}).$$

This means that we consider a weighted function of the total error and the error on the bad loans. It is very important that the model classifies bad loans correctly. This is because people who are incorrectly classified as good cost the financial institution by not receiving payment. A person who is incorrectly classified as bad also costs the financial institution because the financial institution now loses out on payments and thus reduces the institution's profit. The choice of α is subjective depending on how much weight you wish to put on the total error and the error on the bad loans. When $\alpha = 1$ only the total error is minimized and when $\alpha = 0$ only the error on the bad loans is considered. A value between 1 and 0 is suggested so that the total error and the error on the bad loans are both taken into account. The use of this error function results in a more risk averse approach because it results in a lower cut-off probability choice. The error function was calculated on the validation data set when $\alpha = 1$ and $\alpha = 0.8$. Figure 4.15 is obtained when the total error probability is minimized using $\alpha = 1$.

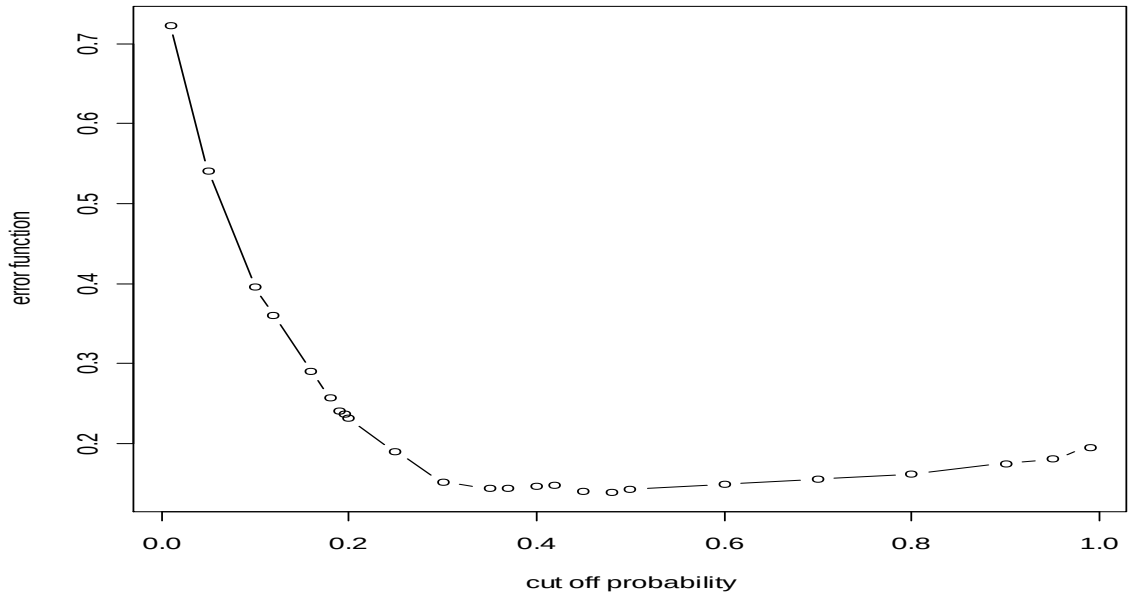


Fig. 4.15 Optimal cut-off probability when total error is minimized.

Whilst if, for example, we use a value of $\alpha = 0.8$, Figure 4.16 is obtained. This value of alpha puts a high weight on the total error while still considering the error on the bad loans. In Figure 4.15, the cut-off probability with the lowest error is 0.48.

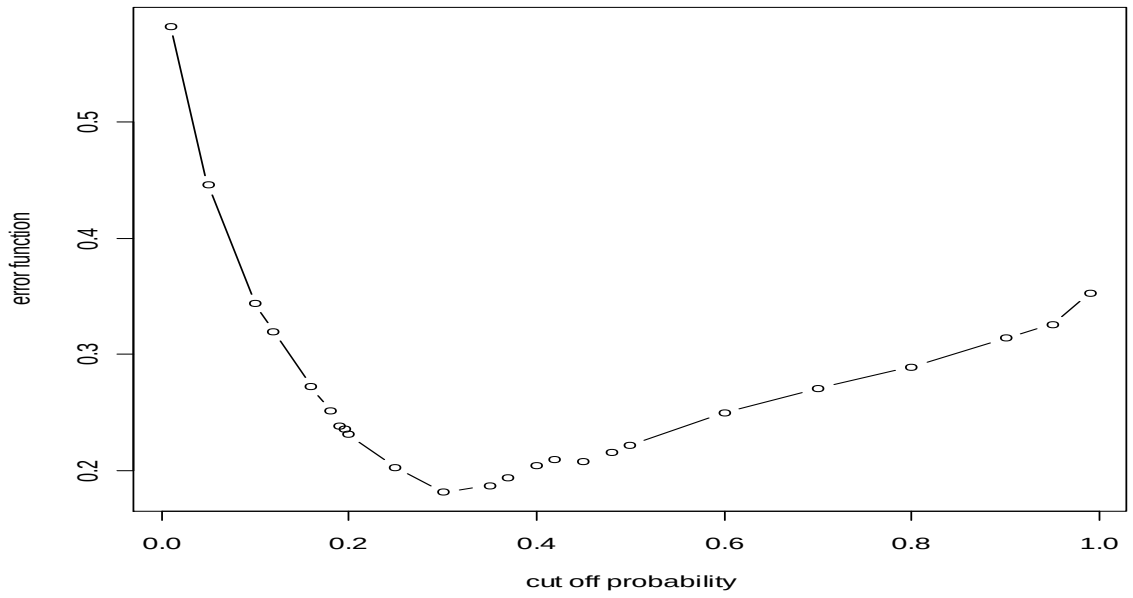


Fig. 4.16 Optimal cut-off probability when error function is minimized with $\alpha = 0.8$.

In Figure 4.16 a cut-off probability of 0.3 gives the lowest value of the error function. This cut-off probability is lower than the cut-off probability when only the total error is minimized. Because we are risk averse, a cut-off probability of 0.3 will be used. This means that anyone with a probability of being bad less than 0.3 will be classified as good, and any with a probability of being bad greater than 0.3 will be classified as bad. Both cut-off probabilities 0.48 and 0.3 were used and the results compared.

4.4 Logistic Regression Model on "new" Data

Six Fisher scoring iterations were needed for the parameters to converge. The estimated parameters are given in Table 4.7.

Table 4.7 Logistic regression model fitted on the "new" data.

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-8.62E+00	1.36E+00	-6.317	2.67E-10	Significant
LOAN	5.85E-06	1.38E-05	0.423	0.67203	Insignificant
MORTDUE	-6.50E-06	6.97E-06	-0.933	0.350944	Insignificant
VALUE	1.62E-06	5.85E-06	0.277	0.781878	Insignificant
REASONHomeImp	1.09E-01	3.22E-01	0.337	0.736028	Insignificant
JOBOffice	-9.80E-01	5.82E-01	-1.684	0.09211	Insignificant
JOBOther	1.62E-01	4.55E-01	0.357	0.721458	Insignificant
JOBProfExe	1.06E-01	5.29E-01	0.2	0.841722	Insignificant
JOBSales	3.33E+00	9.42E-01	3.535	0.000408	Significant
JOBSelf	-1.44E-01	9.04E-01	-0.159	0.873381	Insignificant
YOJ	-2.68E-02	2.15E-02	-1.244	0.213402	Insignificant
DEROG	6.56E-01	2.10E-01	3.125	0.001779	Significant
DELINQ	1.16E+00	1.68E-01	6.904	5.05E-12	Significant
CLAGE	-6.65E-03	2.08E-03	-3.196	0.001394	Significant
NINQ	2.06E-01	6.59E-02	3.122	0.001798	Significant
CLNO	-4.08E-02	1.63E-02	-2.501	0.012374	Significant
DEBTINC	2.33E-01	3.33E-02	6.985	2.86E-12	Significant

What is interesting now is that the variable LOAN has gone from being significant on the "old" data to insignificant on the "new" data, and the JOB variable has a different significant dummy variable. Other than this, the models on the "new" and "old" data are similar. The residual deviance of the model is 341.18 with 549 degrees of freedom.

Interpretation is now given for the parameters of LOAN, DEROG and DEBTINC.

- The parameter of LOAN is 5.85E-06 and is insignificant at the 5% significance level. A unit increase in LOAN with all other variables held fixed, means that there will be a 5.85E-06 increase in the log-odds of default.
- The parameter of DEROG is 6.56E-01 and is significant at the 5% significance level. A unit increase in DEROG with all other variables held fixed, means that there will be a 6.56E-01 increase in the log-odds of default.
- The parameter of DEBTINC is 2.33E-01 and is significant at the 5% significance level. A unit increase in DEBTINC with all other variables held fixed, means that there will be a 2.33E-01 increase in the log-odds of default.

In order to check the adequacy of the model, collinearity of the independent variables, outliers and influential observations are now considered. The correlation matrix of the numerical independent variables is given in Table 4.8.

Table 4.8 Correlation matrix of the numerical independent variables on the "new" data.

Variable	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
LOAN	1.00	0.23	0.33	0.08	-0.02	-0.04	0.10	0.08	0.08	0.07
MORTDUE	0.23	1.00	0.85	0.01	-0.01	-0.02	0.11	0.05	0.31	0.14
VALUE	0.33	0.85	1.00	0.06	-0.04	0.04	0.13	0.04	0.33	0.12
YOJ	0.08	0.01	0.06	1.00	-0.05	0.01	0.16	-0.06	-0.01	-0.09
DEROG	-0.02	-0.01	-0.04	-0.05	1.00	0.14	-0.02	0.07	0.00	0.07
DELINQ	-0.04	-0.02	0.04	0.01	0.14	1.00	0.06	0.02	0.07	0.06
CLAGE	0.10	0.11	0.13	0.16	-0.02	0.06	1.00	-0.1	0.21	-0.07
NINQ	0.08	0.05	0.04	-0.06	0.07	0.02	-0.1	1.00	0.07	0.20
CLNO	0.08	0.31	0.33	-0.01	0.00	0.07	0.21	0.07	1.00	0.17
DEBTINC	0.07	0.14	0.12	-0.09	0.07	0.06	-0.07	0.20	0.17	1.00

From this correlation matrix, we see that there are no large pair-wise correlations. The largest correlation is 0.85 between VALUE and MORTDUE. The other pair-wise correlations are all very small and insignificant. Worrying correlations will occur when the correlation between two variables is greater than 0.9. The variance inflation factors for each numerical variable are given in Table 4.9.

Table 4.9 Variance inflation factors (VIF) of numerical independent variables on the "new" data.

Variable	VIF
LOAN	1.156804
MORTDUE	3.84004
VALUE	4.112173
YOJ	1.047812
DEROG	1.03521
DELINQ	1.046096
CLAGE	1.108899
NINQ	1.063397
CLNO	1.202153
DEBTINC	1.100282

From Table 4.9, there are no large variance inflation factors. Therefore, there is no serious problem with collinearity in the "new" data.

Outliers and influential observations in the model are now considered. A half-normal plot of the residuals is given in Figure 4.17.

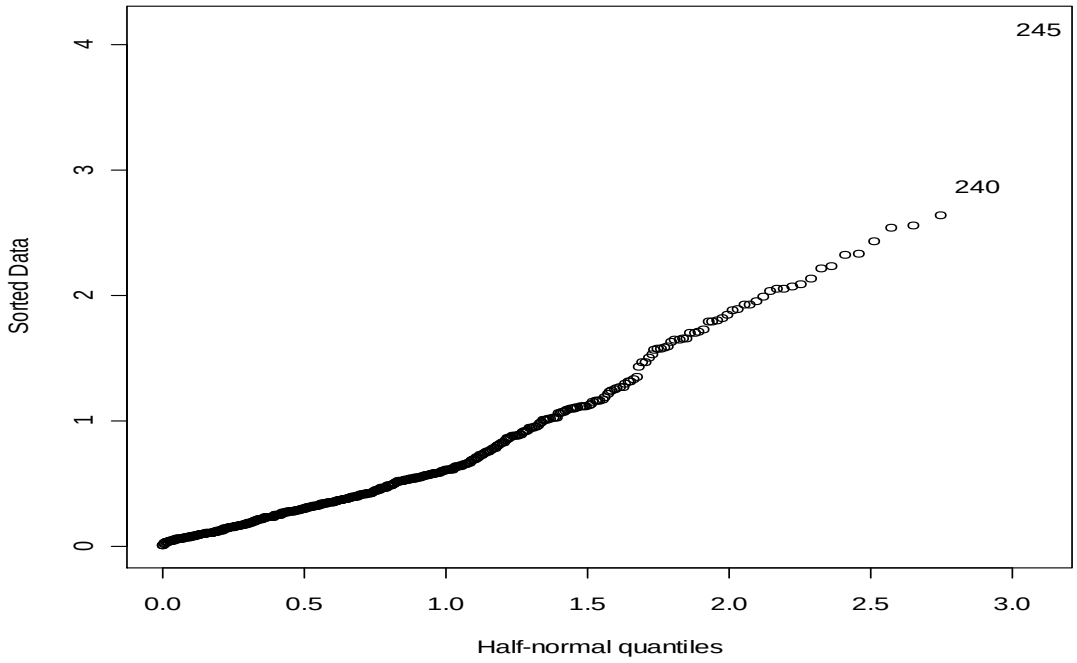


Fig. 4.17 Half-normal plot of residuals for the "new" data.

Figure 4.17 shows no indication of outliers. Secondly, a half-normal plot of the leverages is given in Figure 4.18.

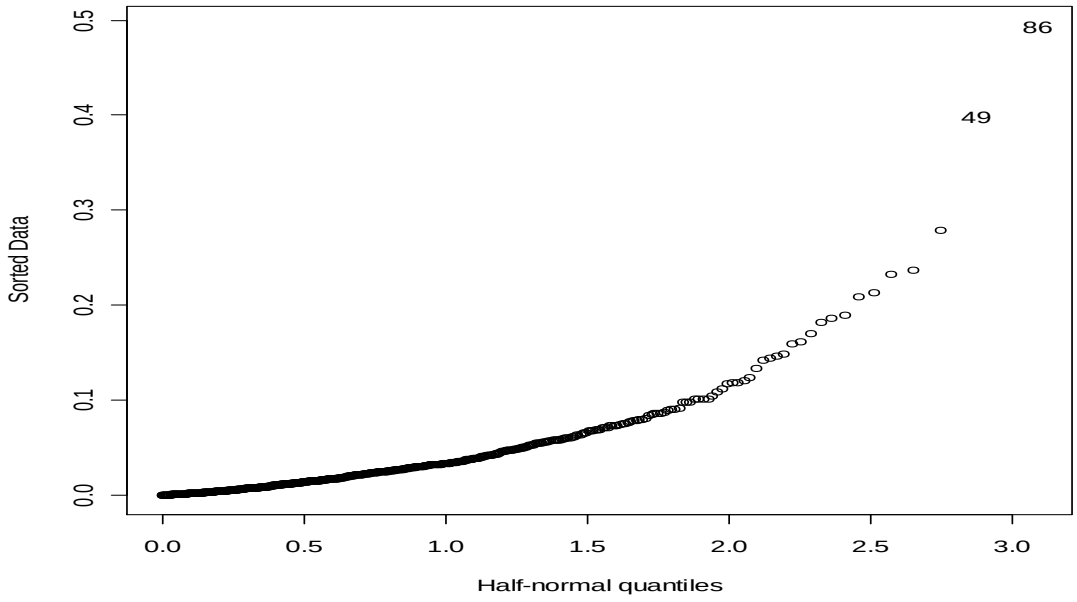


Fig. 4.18 Half-normal plot of leverages for the "new" data.

Figure 4.18 shows that there may be some indication of leverage from observations numbered 49 and 86. Finally a half-normal plot of the Cook's distance statistics is given in Figure 4.19.

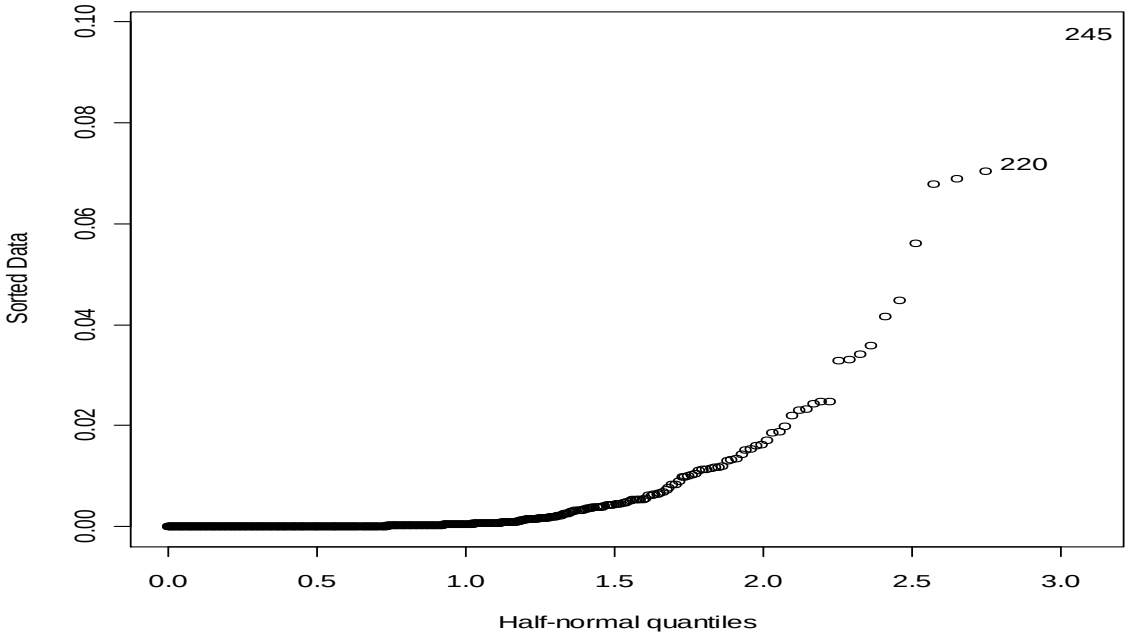


Fig. 4.19 Half-normal plot of the Cook's distance statistics for the "new" data.

From Figure 4.19, there may be some leverage from observations numbered 220 and 245. Thus, observations numbered 49, 86, 220 and 245 may be influential observations. In order to see whether these observations are influential we delete them from the "new" data and re-fit the model. The estimated parameters of this model are then compared to the parameters of the model with no observations deleted. Table 4.10 shows the estimated coefficients.

Table 4.10 Comparison of model coefficients when possible influential observations are either included or excluded from the "new" data.

Variable	Coefficients Including	Coefficents Excluding
(Intercept)	-8.62E+00	-9.97E+00
LOAN	5.85E-06	6.06E-06
MORTDUE	-6.50E-06	-2.02E-06
VALUE	1.62E-06	-1.26E-06
REASONHomeImp	1.09E-01	8.32E-02
JOBOffice	-9.80E-01	-9.88E-01
JOBOther	1.62E-01	2.70E-01
JOBProfExe	1.06E-01	1.54E-02
JOBSales	3.33E+00	4.16E+00
JOBSelf	-1.44E-01	4.59E-02
YOJ	-2.68E-02	-2.70E-02
DEROG	6.56E-01	1.07E+00
DELINQ	1.16E+00	1.21E+00
CLAGE	-6.65E-03	-7.78E-03
NINQ	2.06E-01	1.90E-01
CLNO	-4.08E-02	-4.14E-02
DEBTINC	2.33E-01	2.69E-01

Table 4.10 shows the difference in the parameters when possible influential observations are removed. The first column gives the model parameters with all observations in the "new" data and the second column gives the model parameters when the possible influential observations are removed. Looking at Table 4.10, we see sign changes for the variables VALUE and JOBSelf. Therefore, these observations will be deleted from the "new" data set.

A summary of the model fitted on the data with the influential observations omitted is given in Table 4.11.

Table 4.11 Logistic regression model on "new" data with influential observations removed.

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-9.97E+00	1.51E+00	-6.594	4.28E-11	Significant
LOAN	6.06E-06	1.50E-05	0.404	0.685952	Insignificant
MORTDUE	-2.02E-06	8.22E-06	-0.245	0.806292	Insignificant
VALUE	-1.26E-06	6.93E-06	-0.182	0.855411	Insignificant
REASONHomeImp	8.33E-02	3.38E-01	0.247	0.805237	Insignificant
JOBOffice	-9.88E-01	6.09E-01	-1.623	0.104515	Insignificant
JOBOther	2.70E-01	4.76E-01	0.566	0.571069	Insignificant
JOBProfExe	1.54E-02	5.59E-01	0.027	0.978067	Insignificant
JOBSales	4.16E+00	1.01E+00	4.111	3.94E-05	Significant
JOBSelf	4.60E-02	9.35E-01	0.049	0.960803	Insignificant
YOJ	-2.70E-02	2.24E-02	-1.205	0.228258	Insignificant
DEROG	1.08E+00	3.21E-01	3.348	0.000815	Significant
DELINQ	1.21E+00	1.79E-01	6.786	1.16E-11	Significant
CLAGE	-7.78E-03	2.19E-03	-3.547	0.00039	Significant
NINQ	1.90E-01	6.88E-02	2.762	0.005739	Significant
CLNO	-4.14E-02	1.74E-02	-2.379	0.017338	Significant
DEBTINC	2.69E-01	3.72E-02	7.249	4.20E-13	Significant

The model given in Table 4.11 will be used for prediction. Eight of the 17 variables are significant. The residual deviance of the model is 314.87 with 545 degrees of freedom.

4.5 Bayesian Logistic Regression Model on "new" Data

Now, using the `MCMClogit` function from the `MCMCpack` in `R`, we are able to fit two Bayesian logistic regression models - one with an informative prior and one with a non-informative prior. The MCMC algorithm used is a random walk Metropolis-Hastings algorithm (Section 3.4.3). The Bayesian logistic regression with informative priors will use parameters from the logistic regression on the "old" data as priors. The Bayesian logistic regression model using a non-informative prior will use a uniform prior. The influential observations identified from the logistic regression model on the "new" data were removed. In order to obtain posterior estimates, a Markov chain with 510,000 samples was

generated for both models. The first 500,000 samples were excluded (to allow enough time for the Markov chain to converge to its stationary distribution) which left a Markov chain of 10,000 samples. Therefore, the burn-in period was 500,000.

Bayesian logistic regression model with an informative prior

Prior information came from the model fitted on the "old" data. The model fitted on the "old" data serves as expert information obtained in the home country. This expert knowledge on the logistic regression parameters was then used as prior information for the model on the limited amount of "new" data in the new economic location. A multivariate normal prior is assumed for the parameters. The prior parameters are also assumed to be independent. The prior coefficients are the coefficients from the logistic regression on the "old" data. Each coefficient has corresponding information represented in a 17 x 17 diagonal matrix. The prior coefficients and corresponding element in the diagonal matrix are given in Table 4.12.

Table 4.12 Prior parameters for an informative Bayesian logistic regression model.

Variable	Coefficient	Information
(Intercept)	-7.194241	2.88E+02
LOAN	-2.3673E-05	1.19E+11
MORTDUE	-3.70998E-06	1.94E+12
VALUE	3.03441E-06	3.80E+12
REASONHomeImp	0.2027903	9.37E+01
JOBOffice	-0.681924	3.90E+01
JOBOther	0.01722975	1.36E+02
JOBProfExe	0.04760004	5.46E+01
JOBSales	0.4024014	6.64E+00
JOBSelf	0.4016077	8.87E+00
YOJ	-0.01615048	3.20E+04
DEROG	0.7334939	1.90E+02
DELINQ	0.8039918	3.56E+02
CLAGE	-0.005222989	8.85E+06
NINQ	0.1366665	1.70E+03
CLNO	-0.02814893	1.54E+05
DEBTINC	0.1911389	4.13E+05

In order to get the acceptance rate between 20-40%, a tuning parameter of 0.6 was used. Because a high dimension model is being fitted, the acceptance rate needs to be towards the lower bound of the desired range. The tuning parameter of 0.6 gave an accepted rate of 23%. The model is summarized in Table 4.13.

Table 4.13 Bayesian logistic regression model on the "new" data with an informative prior.

Variable	Mean	SD	2.50%	97.50%
(Intercept)	-7.20E+00	5.52E-02	-7.31E+00	-7.09E+00
LOAN	-2.23E-05	2.90E-06	-2.81E-05	-1.68E-05
MORTDUE	-3.86E-06	6.77E-07	-5.27E-06	-2.57E-06
VALUE	3.03E-06	4.89E-07	2.10E-06	4.03E-06
REASONHomeImp	1.71E-01	1.01E-01	-2.59E-02	3.67E-01
JOBOffice	-7.14E-01	1.48E-01	-1.00E+00	-4.22E-01
JOBOther	3.36E-02	8.64E-02	-1.38E-01	1.98E-01
JOBProfExe	1.51E-02	1.27E-01	-2.19E-01	2.77E-01
JOBSales	8.21E-01	3.48E-01	1.12E-01	1.48E+00
JOBSelf	3.11E-01	3.00E-01	-3.24E-01	8.70E-01
YOJ	-1.58E-02	5.21E-03	-2.56E-02	-4.65E-03
DEROG	7.36E-01	7.20E-02	5.93E-01	8.82E-01
DELINQ	8.35E-01	5.01E-02	7.36E-01	9.38E-01
CLAGE	-5.22E-03	3.21E-04	-5.86E-03	-4.62E-03
NINQ	1.46E-01	2.35E-02	1.00E-01	1.91E-01
CLNO	-2.80E-02	2.57E-03	-3.32E-02	-2.31E-02
DEBTINC	1.92E-01	1.38E-03	1.89E-01	1.94E-01

The mean provides the estimate for the parameter. From Table 4.13, looking at the quantiles for each variable we can determine which variables are significant at the 5% significance level. The values from the 2.5% to the 97.5% quantiles provide a 95% credibility interval for each variable. Only dummy variables for the JOB variable and REASON variable are insignificant. This shows that the majority of variables included in the model are significant in predicting good and bad applicants. The parameter estimates

still have the same interpretation and interpretations for the parameters of LOAN, DEROG and DEBTINC are:

- The parameter of LOAN is -2.23E-05 and is significant at the 5% significance level. The reason for this is that the 95% credibility interval does not contain zero. A unit increase in LOAN with all other variables held fixed, means that there will be 2.23E-05 decrease in the log-odds of default.
- The parameter of DEROG is 7.36E-01 and is significant at the 5% significance level since its credibility interval does not contain zero. A unit increase in DEROG with all other variables held fixed, means that there will be a 7.36E-01 increase in the log-odds of default.
- The parameter of DEBTINC is 1.92E-01 and is significant at the 5% significance level because its credibility interval does not contain zero. A unit increase in DEBTINC with all other variables held fixed, means that there will be a 1.92E-01 increase in the log-odds of default.

Trace plots of the Markov chain and density plots of the posterior distributions are given in Figure 4.20. Trace and density plots are only given for the first four parameters. The remaining trace and density plots can be found in the Appendix D1.

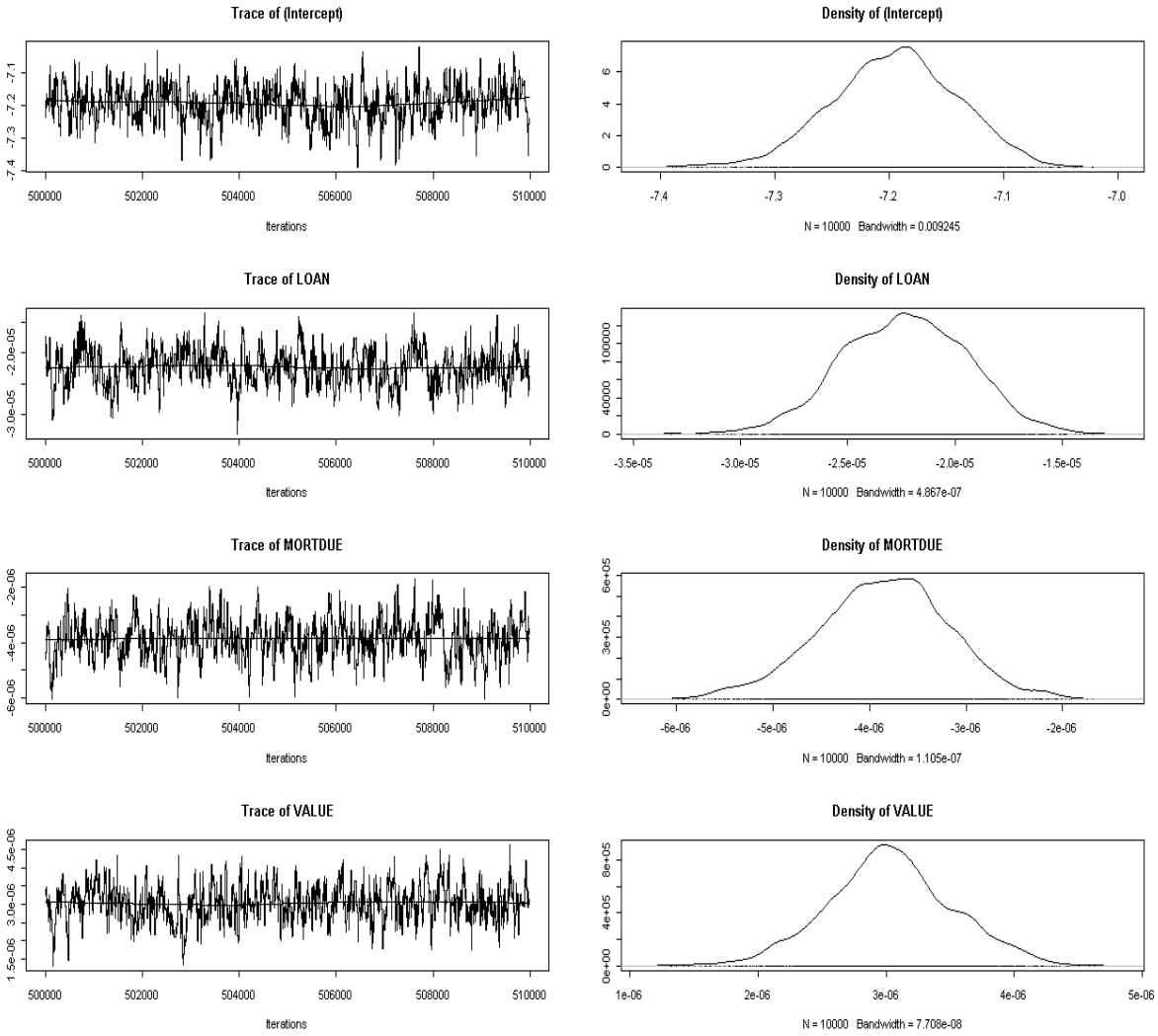


Fig. 4.20 Trace and density plots of the posteriors for the first four variables using an informative prior.

From Figure 4.20, looking at the trace plot of the Markov chain, the Markov chain is relatively stationary. This implies that the Markov chain has reached or is close to its stationary distribution. A concern is that the Markov chain still appears to be quite strongly correlated. The density plots of the first four variables show an irregular bell shaped distribution. The remaining trace and density plots are similar (Appendix D1).

The Geweke diagnostic statistics are now calculated for each variable and are given in Table 4.14.

Table 4.14 Geweke diagnostic statistics for each variable of the Bayesian logistic regression model with informative prior.

Variable	z
(Intercept)	1.3294
LOAN	0.8905
MORTDUE	-0.7683
VALUE	-0.5705
REASONHomeImp	-0.4904
JOBOffice	1.9190
JOBOther	0.8031
JOBProfExe	-0.5658
JOBSales	0.4618
JOBSelf	-2.1442
YOJ	-2.5343
DEROG	0.6239
DELINQ	1.8627
CLAGE	0.4984
NINQ	0.9205
CLNO	-1.6003
DEBTINC	-0.6575

Table 4.14 shows that the variables YOJ and JOBSelf have $|z| > 2$. Therefore, these two variables have not converged. All the other variables have converged according to the Geweke diagnostic.

Bayesian logistic regression model with non-informative prior

An improper uniform prior is now used as a prior. This is a non-informative prior and provides no prior information for the parameters. Again a tuning parameter of 0.6 was used in order to get the acceptance rate in the lower end of the 20-40% range. The acceptance rate was 24.5%. The model is summarized in Table 4.15.

Table 4.15 Bayesian logistic regression model with non-informative prior on the "new" data.

Variable	Mean	SD	2.50%	97.50%
(Intercept)	-9.21E+00	1.54E+00	-1.24E+01	-6.47E+00
LOAN	5.25E-06	1.40E-05	-2.32E-05	3.12E-05
MORTDUE	-7.72E-06	6.98E-06	-2.20E-05	5.29E-06
VALUE	2.17E-06	5.73E-06	-9.06E-06	1.40E-05
REASONHomeImp	9.08E-02	3.56E-01	-5.87E-01	7.70E-01
JOBOffice	-9.89E-01	5.95E-01	-2.10E+00	1.34E-01
JOBOther	1.65E-01	4.72E-01	-7.11E-01	1.12E+00
JOBProfExe	1.18E-01	5.56E-01	-9.44E-01	1.28E+00
JOBSales	3.57E+00	9.67E-01	1.64E+00	5.49E+00
JOBSelf	-2.27E-01	9.68E-01	-2.18E+00	1.62E+00
YOJ	-2.85E-02	2.23E-02	-7.15E-02	1.42E-02
DEROG	7.69E-01	2.21E-01	3.74E-01	1.27E+00
DELINQ	1.24E+00	1.74E-01	8.96E-01	1.58E+00
CLAGE	-7.20E-03	2.08E-03	-1.11E-02	-3.07E-03
NINQ	2.12E-01	7.22E-02	6.88E-02	3.51E-01
CLNO	-4.49E-02	1.74E-02	-7.77E-02	-1.06E-02
DEBTINC	2.52E-01	3.64E-02	1.89E-01	3.30E-01

The parameter estimates still have the same interpretation and interpretations for the parameters of LOAN, DEROG and DEBTINC are:

- The parameter of LOAN is 5.25E-06 and is insignificant at the 5% significance level. The reason for this is that the 95% credibility interval does contain zero. A unit increase in LOAN with all other variables held fixed, means that there will be a 5.25E-06 increase in the log-odds of default.
- The parameter of DEROG is 7.69E-01 and is significant at the 5% significance level since its credibility interval does not contain zero. A unit increase in DEROG with all other variables held fixed, means that there will be a 7.69E-01 increase in the log-odds of default.
- The parameter of DEBTINC is 2.52E-01 and is significant at the 5% significance level because its credibility interval does not contain zero. A unit increase in DEBTINC with all

other variables held fixed, means that there will be a 2.52E-01 increase in the log-odds of default.

Trace plots of the Markov chain and density plots of the posterior distributions are given in Figure 4.21. Trace and density plots are only given for the first four parameters. The remaining trace and density plots can be found in Appendix D2.

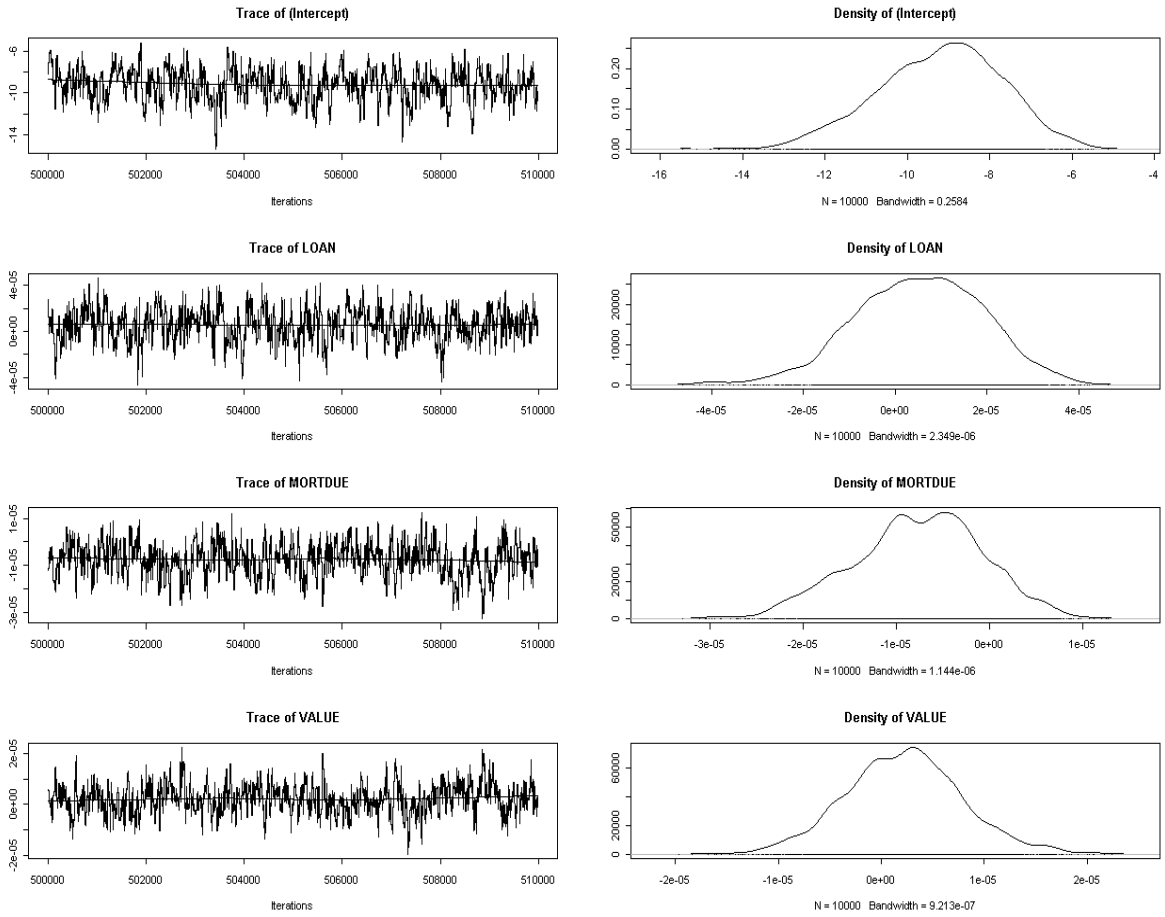


Fig. 4.21 Trace and density plots of the posteriors for the first four variables using a non-informative prior.

From Figure 4.21, looking at the trace plot of the Markov chain, the Markov chain is relatively stationary. This implies that the Markov chain has reached or is close to its stationary distribution. The density plots appear to be bell shaped. The remaining trace and density plots are similar (Appendix D2).

The Geweke diagnostic statistics for each variable are given in Table 4.16.

Table 4.16 Geweke test statistics for each variable for the Bayesian logistic regression model with non-informative prior.

Variable	z
(Intercept)	2.075362
LOAN	0.43113
MORTDUE	0.260332
VALUE R	-0.002344
EASONHomeImp	-1.250048
JOBOffice	-0.715934
JOBOther	-1.383707
JOBProfExe	-1.587358
JOBSales	-2.926755
JOBSelf	-0.688299
YOJ	-0.351282
DEROG	0.056095
DELINQ	2.13558
CLAGE	-0.887012
NINQ	-0.095634
CLNO	-1.234432
DEBTINC	-1.444675

Table 4.16 shows that the intercept, JOBSales and DELINQ have $|z| > 2$ and, therefore, have not converged. All the other variables have converged according to the Geweke diagnostic.

4.6 Performance of Models on Test Data

A test set is now used to assess the performance of the three models on the "new" data: the logistic regression model (Model 1), the Bayesian logistic regression model with informative prior (Model 2) and the Bayesian logistic regression model with non-informative prior (Model 3). The test set contains 1,662 observations and is assumed to come from the new economic location. The performance of the models on the test set was compared for two cut-off probabilities, 0.3 and 0.48.

4.6.1 Cut-off probability of 0.3

Logistic regression model

Table 4.17 Classification table for the logistic regression model with cut-off probability of 0.3.

		Predicted	
		Good	Bad
Actual	Good	1126	196
	Bad	100	240

Of the 1,662 applicants in the test set, the logistic regression model (Model 1) rejected 436 and accepted 1,226 applicants. Of the rejected applicants, 196 (45.0%) are in fact good (Table 4.17). Therefore, 196 applicants are missed profits for the financial institution. Of the accepted applicants, 100 (8.2%) were bad - losses for the financial institution. Because the financial institution is only exposed to the applicants it accepted, the classification error is 8.2%. The overall classification error rate is 17.8%. The overall classification error rate gives a better indication since it includes applicants which represent missed out profits for the financial institution.

Bayesian logistic regression model with an informative prior

Table 4.18 Classification table for the Bayesian logistic regression model with informative prior and cut-off probability of 0.3.

		Predicted	
		Good	Bad
Actual	Good	1155	167
	Bad	106	234

Of the 1,662 applicants in the test set, the Bayesian logistic regression model with informative prior (Model 2) rejected 401 and accepted 1,261 applicants (Table 4.18). Of the rejected applicants, 167 (41.6%) are in fact good - this is missed out profits for the

financial institution. Of the accepted applicants, 106 (8.4%) were bad. This represents losses for the financial institution. The classification error rate realized by the financial institution is thus 8.4%. The overall classification error rate is 16.4%.

Bayesian logistic regression model with a non-informative prior

Table 4.19 Classification table of the Bayesian logistic regression model with non-informative prior and cut-off probability of 0.3.

		Predicted	
		Good	Bad
Actual	Good	1141	181
	Bad	101	239

Of the 1,662 applicants on the test set, the Bayesian logistic regression model with non-informative prior (Model 3) rejected 420 and accepted 1,242 applicants (Table 4.19). Of the rejected applicants, 181 (43.1%) are in fact good - missed out profits for the financial institution. Of the accepted applicants, 101 (8.1%) were bad - losses for the financial institution. 8.1% is thus the classification error rate realized by the financial institution. The overall classification error rate is 17.0%.

Comparison of the 3 models

Table 4.20 compares Models 1, 2 and 3.

Table 4.20 Comparison of Models 1, 2 and 3 when the cut-off probability is 0.3.

	Model 1	**Model 2**	**Model 3**
Accepted	1226	1261	1242
Rejected	436	401	420
Error rate among accepted	8.2%	8.4%	8.1%
Error rate among rejected	45.0%	41.6%	43.1%
Total error rate	17.8%	16.4%	17.0%

From Table 4.20, the following can be deduced:

- Model 2 accepts the most applicants.
- Model 1 rejects the most applicants.
- Model 3 has the lowest error rate among the accepted applicants.
- Model 2 has the lowest error rate among the rejected applicants.
- Model 2 has the lowest total error rate.

In terms of total error rate, the best model is Model 2 and the second best model is Model 3. Therefore, both the Bayesian models perform better than the logistic regression model. For the error rates among the accepted applicants (the error realized by the financial institution), the error rates are fairly close to each other. Model 2 is thus the best model to use as it would result in the most profit for the financial institution.

4.6.2 Cut-off probability of 0.48

For a comparison, if a cut-off probability was chosen to minimize the total error rate on the validation set, the cut-off probability is 0.48. Using this cut-off probability would mean more risk for the financial institution. When 0.48 is used as a cut-off the following results are obtained.

Logistic regression model

Table 4.21 Classification table of logistic regression model with cut-off probability of 0.48.

		Predicted	
		Good	Bad
Actual	Good	1215	107
	Bad	137	203

Of the 1,662 applicants in the test set, the logistic regression model (Model 1) now rejects 310 and accepts 1,352 applicants (Table 4.21). 107 (34.5%) or the rejected applicants are in fact good. 137 (10.1%) of the accepted applicants are bad. The overall classification error rate is 14.7%.

Bayesian logistic regression model with informative prior

Table 4.22 Classification table of Bayesian logistic regression model with informative prior and cut-off probability of 0.48.

		Predicted	
		Good	Bad
Actual	Good	1267	55
	Bad	152	188

Of the 1,662 applicants in the test set, the Bayesian logistic regression model with informative prior (Model 2) now rejects 243 and accepts 1,419 applicants (Table 4.22). 55 (22.6%) of the rejected applicants are in fact good. 152 (10.7%) of the accepted applicants are bad. The overall classification error rate is 12.5%.

Bayesian logistic regression model with a non-informative prior

Table 4.23 Classification table of the Bayesian logistic regression model with non-informative prior and cut-off probability of 0.48.

		Predicted	
		Good	Bad
Actual	Good	1237	85
	Bad	143	197

Of the 1,662 applicants in the test set, the Bayesian logistic regression model with non-informative prior (Model 3) now rejects 282 and accepts 1,380 applicants (Table 4.23). 85 (30.1%) of the rejected applicants are in fact good. 143 (10.4%) of the accepted applicants are bad. The overall classification error rate is 13.7%.

Comparison of the 3 models

Table 4.24 compares Models 1, 2 and 3.

Table 4.24 Comparison of Models 1, 2 and 3 when the cut-off probability is 0.48.

	Model 1	**Model 2**	**Model 3**
Accepted	1352	1419	1380
Rejected	310	243	282
Error rate among accepted	10.1%	10.7%	10.4%
Error rate among rejected	34.5%	22.6%	30.1%
Total error rate	14.7%	12.5%	13.7%

The following can be deduced from Table 4.24:

- Model 2 accepts the most applicants.
- Model 1 rejects the most applicants.
- Model 1 has the lowest error rate among the accepted applicants.
- Model 2 has the lowest error rate among the rejected applicants.
- Model 2 has the lowest total error rate.

The error rates amongst the accepted applicants are all fairly close for all the models. The Bayesian logistic regression model with informative prior again has the lowest total error rate, showing that the use of relevant prior information is beneficial.

4.6.3 Comparison of the two cut-off probabilities

The results are now compared when the two different cut-off probabilities are used, 0.3 and 0.48. The following conclusions are reached:

- In all models, more applicants are accepted when the cut-off probability is 0.48 as opposed to 0.3.
- In all models, fewer applicants are rejected when the cut-off probability is 0.48 as opposed to 0.3.
- For all models, the error rate among the accepted applicants is higher when the cut-off probability is 0.48 as opposed to 0.3. This shows that the error rate realized by the bank is lower when a lower cut-off probability is used.
- For all models, the error rate among the rejected applicants is lower with the cut-off probability is 0.48 as opposed to 0.3.
- For all models, the total error rate is lower when the cut-off probability is 0.48 as opposed to 0.3.

It appears that 0.48 is a better cut-off probability to use because it exposes the financial institution to more people who will be good. This means that the financial institution will be more profitable than one which uses a cut-off probability of 0.3. The difference in the error rates among the accepted applicants for the two cut-off probabilities is around 2% for each model. This is not big enough to opt for the conservative approach of a cut-off probability of 0.3. The financial institution may want to employ the more risk averse cut-off probability if it expects the financial markets to become turbulent.

4.7 Performance of the Models with Varying Amounts of "new" Data

The performance of Models 1, 2 and 3 were then compared when the amount of "new" data varied. Again two different cut-off probabilities were used, namely 0.3 and 0.48.

4.7.1 Cut-off probability of 0.3

The error rates of Models 1, 2 and 3 when the models are trained using differerent sampler sizes and the cut-off probability is 0.3 are given in Figure 4.22.

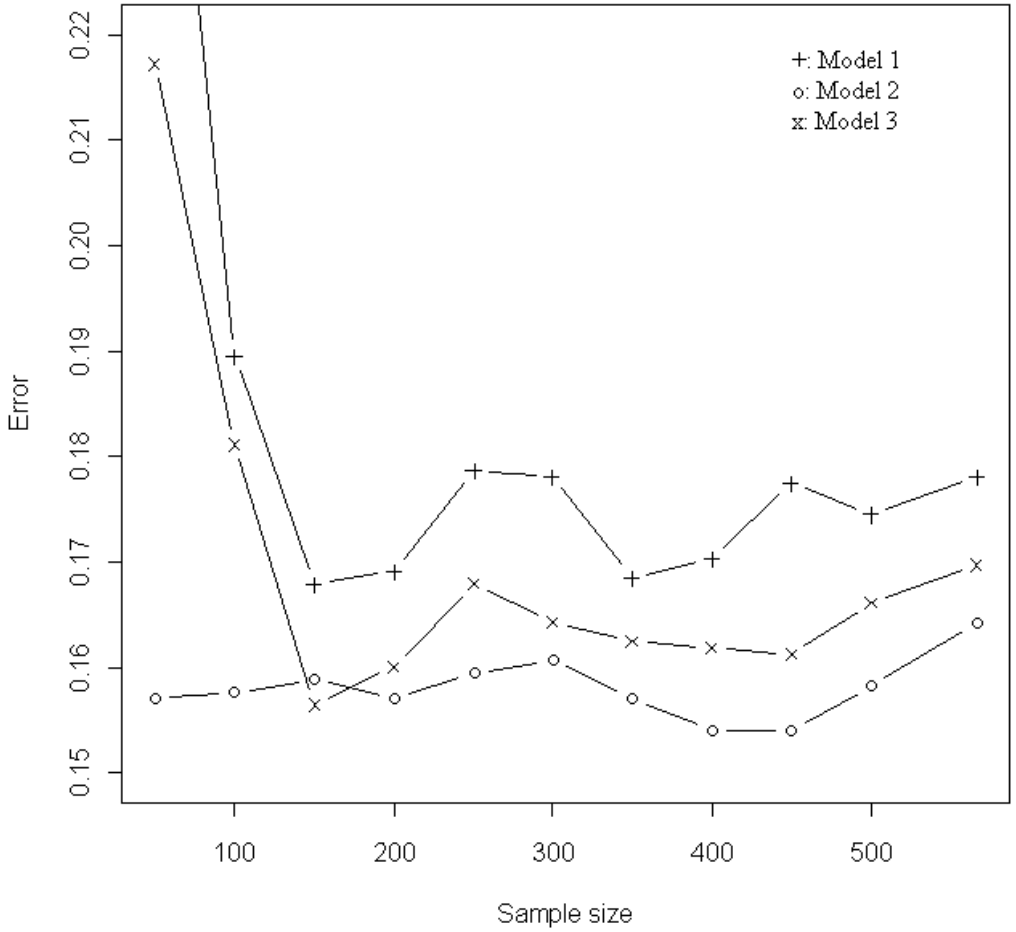


Fig. 4.22 Error rates of Models 1, 2 and 3 when the models are trained using differerent sampler sizes and the cut-off probability is 0.3.

From Figure 4.22, Models 1 and 3 appear to follow a similar pattern. The error rate of Model 3 is always below that of Model 1. The error rates of Models 1 and 3 appear to decrease as the sample size of the "new" data increases. The Bayesian model with non-informative prior thus appears to perform better than the logistic regression model. The

error rate of Model 2 is relatively stable. The error rate of this model is always below the other two models (except when the sample size is 150, Model 3 has a lower error). This shows that making use of prior information is very useful when the sample size is small. It is expected that the error rates of these three models would converge as the sample size increases.

When a financial institution is expanding into a new economic location, combining expert information obtained from experience in the home country can be very useful.

4.7.2 Cut-off probability of 0.48

For a comparison, if a cut-off probability was chosen to minimize the total error rate on the validation set, the cut-off probability would have been 0.48. Using this cut-off probability would mean more risk for the financial institution. When 0.48 is used as a cut-off probability Figure 4.23 is obtained.

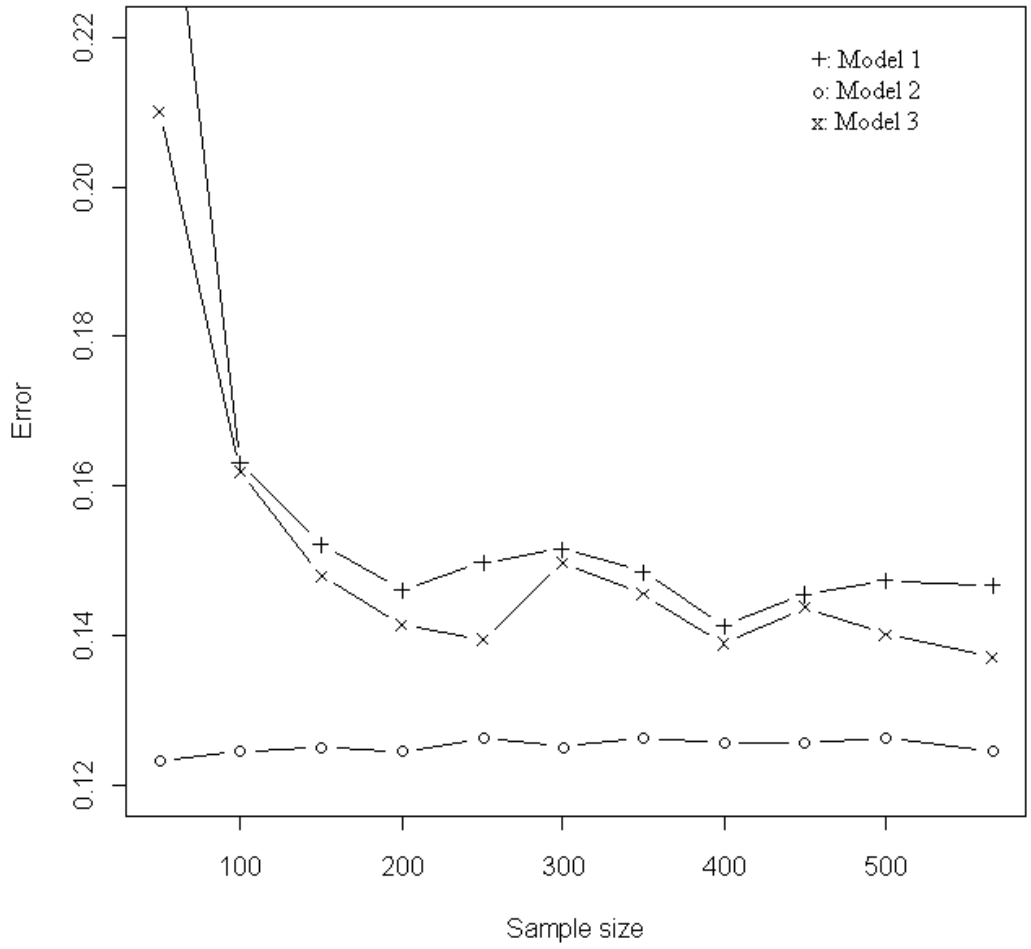


Fig. 4.23 Error rates of Models 1, 2 and 3 when the models are trained using differerent sampler sizes and the cut-off probability is 0.48.

Figure 4.23 again shows Models 1 and 3 following a similar pattern. The trend of the error rate decreasing as the sample size of the "new" data increases is also clear. The error rates of Models 1 and 3 decrease as the sample size increases but then appear to level-off. The error rate of Model 2 is again relatively constant. This graph confirms again that the use of relevant prior information is very useful.

The total error rates when the cut-off probability is 0.48 are lower than when the cut-off probability is 0.3 (Figures 4.22 and 4.23). Model 2 performs better when a cut-off probability of 0.48 is used.

Although the models perform better on the total error rate with a cut-off probability of 0.48, the error rate amongst the accepted applicants is still higher than when a cut-off probability of 0.3 is used. This confirms that opting for a higher cut-off probability results in accepting more applicants and thus results in more profits. It appears that it is better to take on more risk and opt for a cut-off probability of 0.48.

4.8 Conclusions

This chapter discussed the results obtained by fitting the relevant models on the data. The initial data analysis showed that there are a large number of missing values and that these missing values needed to be estimated in order to keep the proportion of "bads" in the data from dropping. The initial data analysis also showed that there may be a problem with outliers and influential observations.

The logistic regression model fitted on the "old" data was then used to determine cut-off probabilities. Two potential cut-off probabilities were obtained: one where the total error rate was minimized and one where a weighted error function was minimized taking into account the error rate on the "bads". The cut-off probability obtained when the total error rate was minimized was 0.48. The minimization of the weighted error function resulted in a more conservative cut-off probability of 0.3.

A logistic regression model (Model 1), a Bayesian logistic regression with priors from the logistic regression on the "old" data (Model 2), and a Bayesian logistic regression model with non-informative priors (Model 3) were then fitted on the "new" data. All the fitted models had predictor variables with a number of significant parameters.

The performance of these models was then compared on a test set. With cut-off probabilities of 0.3 and 0.48, Model 2 performs the best in terms of total error rate. It was found that with a higher cut-off probability, more people were accepted which results is a slightly higher error rate among the accepted applicants. The error rates realized by the financial institution (error rate on accepted applicants) were higher for all 3 models when a

higher cut-off probability was used. The total error rates for all three models were lower when the higher cut-off probability was used. This shows that it is worth taking on more risk to realize more profit.

When the sample size of the "new" data set was varied for the training of the models, it was found that Model 2 performed the best. Model 3 was also found to perform better than Model 1. Models 1 and 3 showed a similar pattern with the error rate decreasing as the sample size of the "new" data set increased. The error rates for Models 1 and 3 appear to level-off at a certain point, whilst Model 2 appears to have a relatively constant error rate for the varying sample sizes.

Making use of relevant prior information can, therefore, be very important in improving the accuracy of credit scoring models. This prior information is most useful when the size of the data set available is small. The importance of the prior information decreases quickly as the amount of available data increases. When there is a lot of data available there is no need to conduct a Bayesian logistic regression- a standard logistic regression will give very similar results.

The following recommendation is, therefore, made to a financial institution expanding into a new economic location: making use of relevant prior information obtained from experiences gained in the home country or other countries can be very useful initially. As the amount of data increases in the new country the usefulness of this prior information decreases and becomes less important.

Chapter 5: Conclusions and Implications

5.1 Summary

This study provided an investigation into the use of Bayesian logistic regression models for credit scoring. The main aim was to determine whether the use of relevant prior information was useful for a financial institution when it was having data quantity issues.

The first step of the study was to review existing literature. It was found that there are a number of models which can be used to build a credit scoring model - there is, however, no "best" model. Bayesian logistic regression models with relevant prior information were shown to provide an improvement over other models when the amount of data available to train the model is small. The literature review was then followed by a theory section. This section provided the theory which was used in the study.

In the results chapter, models were fitted to the data. The data set was randomly split into four sets, *viz*. the "old", "new", validation and test data sets. For a financial institution expanding into a new economic location, the "old" data were assumed to come from the home location, the "new" data from the new location, the validation data from the old location and the test data from the new location. The financial institution was looking to build a scoring model in the new economic location with a limited amount of data. A logistic regression model was fitted on the "old" data and the parameters from this model were used as prior information for a Bayesian logistic regression model with informative prior in the "new" data. A logistic regression and Bayesian logistic regression with non-informative prior was also fitted on the "new" data. Using two different cut-off probabilities for classification, it was found that on the test set, the Bayesian logistic regression model with informative prior provided a lower total error rate.

The error rates of the models were also compared when there are different amounts of "new" data available to build the model. It was found that the Bayesian logistic regression with informative prior provided relatively constant error rates. The logistic regression model and Bayesian logistic regression with non-informative prior had error rates which

started high when the amount of "new" data was small, and subsequently as the size of the "new" data increased, these error rates decreased and levelled off. The Bayesian logistic regression with non-informative prior gave lower error rates than the logistic regression model. The pattern of the error rates as the amount of data increased for these two models was, however, similar.

The use of prior information is very useful when a financial institution is expanding into a new economic location and at first has limited data available. The usefulness of using relevant prior information decreases as the amount of data available increases.

5.2 Limitations, Recommendations and Further Research

A limitation of this study was that the same data set was used to obtain prior information. The prior information used, therefore, reflects a situation when there is "perfect prior information". Although this is a limitation, in practice there will be experts with much experience in credit scoring who would be able to provide very good prior information. A way to perhaps solve this problem would be to change the structure of the "old" and "new" data. This could be done by removing variables from the "old" data. There would then be only prior information for the variables in the "old" data. The usefulness of this reduced prior information could be explored.

It is very difficult to obtain credit scoring data from financial institutions. However, if one could get hold of two data sets from a financial institution, one from its home economic location and one from a new economic location, a more realistic analysis could be done. If these data could be obtained, a better insight would be gained into the use of Bayesian logistic regression models in practice.

Only one method to estimate the missing values in the data set was used. There are a number of methods which are available for the estimation of missing values. Simply replacing the missing values by the overall mean for each variable and the EM (Expectation-Maximization) algorithm are other possible methods which could be

investigated. This latter method works by substituting values iteratively in conjunction with a model. A comparison of these different estimation methods in credit scoring is another area for further research.

This study considered normally distributed priors for the Bayesian models with informative priors. Priors with other distributions can also be considered, for example, the beta distribution. The Laplace prior is another area for future research. For Bayesian models with non-informative priors, an improper uniform prior was used in this study. There are other possible choices, namely the Jeffreys' non-informative priors.

There appeared to be a minor issue with the convergence of the MCMC algorithms. From the trace plots, there was possibly some significant autocorrelation in the Markov chain. Methods to remove this correlation could also be considered - such as thinning. Thinning reduces the sample size of the generated Markov chain by only taking every 2nd or 3rd observation. The Geweke diagnostic also showed that the generated Markov chain for some of the variables had not converged. The analysis could be done again this time using a larger burn-in period for the generated Markov chains (to allow for more time for the chain to converge). This could result in better parameter estimates for the Bayesian models.

This study used a random-walk Metropolis-Hastings algorithm to sample from the posterior distributions. There are a number of algorithms available. The independence sampler is another method which could be considered. The use of a Gibbs sampler when auxiliary variables are used is another interesting model which could be investigated.

There are many different models which can be used to build a credit scoring model. One method which has shown some success in a Bayesian framework is Bayesian networks as shown by Biçer *et al.* (2010). An investigation into these networks for credit scoring would also provide interesting further topics of research.

References

Altman, E., Marco, G., and Varetto, F. (1994). Corporate distress diagnostics: Comparison using linear discriminant analysis and neural networks (the Italian Experience). *Journal of Banking and Finance* 18: 505-529.

Bernardo, J.M. and Smith, A.F.M. (2000). *Bayesian Theory*. John Wiley & Sons, Chichester.

Biçer, I., Sevis, D., and Bilgiç, T. (2010). Bayesian credit scoring model with integration of expert knowledge and customer data. *Twenty-fourth Mini EURO Conference on Continuous Optimization and Information-Based Technologies in the Financial Sector*, Vilnius Gediminas Technical University Publishing House, Technika, pp. 324–329.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* 49(4): 327-335.

Crook, J. and Banasik, J. (2002). Does reject inference really improve the performance of application scoring models? *Working Paper Series* No. 02/3. The Credit Research Centre, The School of Management, University of Edinburgh, pp. 1-27.

Desai, V.S., Crook, J.N., and Overstreet, G. (1996). Credit scoring models in credit union environment. *European Journal of Operational Research* 95: 24-35.

Dobson, A.J. and Barnett, A.G. (2008). *An Introduction to Generalized Linear Models.* 3rd Edition. Taylor & Francis Group, Boca Raton, Florida.

Durand, D. (1941). Credit-rating formulae. In: *Risk Elements in Consumer Instalment Financing*. pp. 83-91. National Bureau of Economic Research, Inc. Massachusetts.

Faraway, J.J. (2006). *Extending the Linear Model with R.* Taylor & Francis Group, Boca Raton, Florida.

Fernandes, G. and Rocha, C.A. (2011). *Low Default Modelling*: *A Comparison of Techniques Based on a Real Brazilian Corporate Portfolio.* Available from

http://www.crc.man.ed.ac.uk/conference/archive/2011/Fernandes-Guilherme-Paper-Low-default-modelling.pdf. Accessed date: 8th November 2011.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics* 7(2): 179-188.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85: 398-409.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: *Bayesian Statistics* 4, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F M. Smith, pp. 169-194, Clarendon Press, Oxford.

Greenberg, E. (2008). *Introduction to Bayesian Econometrics*. Cambridge University Press, New York.

Hand, D.J. and Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*, Series A 160: 523-541.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika* 57: 97–109.

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1: 145-168.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* 2nd Edition. John Wiley & Sons, Inc., New York.

Komorád, K. (2002). *On Credit Scoring Estimation.* Master's Thesis. Humboldt University, Berlin. Available from http://edoc.hu-berlin.de/master/komorad-karel-2002-12-18/PDF/komorad.pdf. Accessed date: 10th January 2011.

Kroese, D.P., Taimre, T., and Botev, Z.I. (2011). *Handbook of Monte Carlo Methods.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Lee, P.M. (2004). *Bayesian Statistics*, *An Introduction.* 3rd Edition. Hodder Arnold, London.

Löffler, G., Posch, P.N., and Schöne, C. (2005). Bayesian methods for improving credit scoring models. *Technical report, Department of Finance, University of Ulm*, Germany.

Mendenhall, W. and Sincich, T. (2003). *A Second Course in Statistics Regression Analysis.* 6th Edition. Pearson Education, Inc, New Jersey.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21: 1087-1092.

Mira, A. and Tenconi, P. (2004). Bayesian estimate of credit risk via MCMC with delayed rejection. In: *Seminar on Stochastic Analysis, Random Fields and Applications IV*. Centro Stefano Franscini, Ascona, pp. 277-291. Birkhauser Verlag, Basel.

Mok, J-M. (2009). *Reject Inference in Credit Scoring.* Available from http://www.few.vu.nl/en/Images/werkstuk-mok_tcm39-91398.pdf. Accessed date: 5th March 2011.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A 132: 370-384.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Press, S.J. (1989). *Bayesian Statistics*: *Principles, Models, and Applications*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* 2nd Edition. Springer-Verlag, New York.

Robert, C.P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R.* Springer-Verlag, New York.

Rona-Tas, A. and Hiß, S. (2008). *Consumer and Corporate Credit Ratings and the Subprime Crisis in the U.S. with some Lessons for Germany*. Available from http://weber.ucsd.edu/~aronatas/The%20Subprime%20Crisis%202008%2010%2004.pdf. Accessed date: 22nd September 2011.

Steenackers, A. and Goovaerts, M. (1989). A credit scoring model for personal loans. Insurance: *Mathematics and Economics* 8: 31-34.

Suess, E.A. and Trumbo, B.E. (2010). *Introduction to Probability Simulation and Gibbs Sampling with R.* Springer-Verlag, New York.

Thomas, L.C. (2009). *Consumer Credit Models*: *Pricing, Profit and Portfolios.* Oxford University Press, Oxford.

Whittacker, J., Whitehead, C., and Somers, M. (2007). A dynamic mortgage scorecard using Kalman filtering. *Journal of the Operational Research Society* 58: 911-921.

Wielenga, D., Lucas, B., and Georges, J. (1999). *Enterprise Miner*^{TM}: *Applying Data Mining Techniques Course Notes*. SAS Institute Inc., Cary, North Carolina.

Wilhelmsen, M., Dimakos, X.K., Husebø, T., and Fiskaaen, M. (2009). *Bayesian Modelling of Credit Risk using Integrated Nested Laplace Approximations.* Available from http://publications.nr.no/BayesianCreditRiskUsingINLA.pdf. Accessed date: 21st November 2010.

Wood, S.N. (2006). *Generalized Additive Models, An Introduction with R*. Taylor & Francis Group, Boca Raton, Florida.

Ziemba, A. (2005). *Bayesian Updating of Generic Scoring Models.* Available from http://www.crc.man.ed.ac.uk/conference/archive/2005/papers/ziemba-arkadius.pdf. Accessed date: 25th November 2010.

Appendix

Appendix A: Code

```
library(faraway)

old_data=na.omit(old_data)        #removing missing values in categorical
variables

new_data=na.omit(new_data)

val_data=na.omit(val_data)

test_data=na.omit(test_data)

##############################################################################
##

mod_old=glm(BAD~.,family=binomial,old_data)    #logit model on old data

mod_new=glm(BAD~.,family=binomial,new_data)    #logit model on new data

##############################################################################
##

x0=val_data[,2:13]  #selecting independent variables

ps=c(0.01,0.05,0.1,0.12,0.16,0.18,0.19,0.195,0.2,0.25,0.3,0.35,0.37,0.4,0.42,0.45
,0.48,0.5,0.6,0.7,0.8,0.9,0.95,0.99)    #cut-off probabilities

error_function=0                  #initialize error function

for(i in 1:length(ps)){

y=ilogit(predict(mod_old,x0))

y[y>ps[i]]<-1

y[y<=ps[i]]<-0

table=table(val_data$BAD,y)

error_function[i]=0.8*((table[2]+table[3])/sum(table))+0.2*(table[2]/(table[2]+ta
ble[4]))

}                   #determining which cut-off gives lowest error


plot(ps,error_function,type="b",xlab="cut-off probability",ylab="error function")

#############################################################################

#checking assumptions on old data

#collinearity
```

```
x1=model.matrix(mod_old)[,-1]

x1=x1[,c(-4:-9)]           #remove categorical variables

cor(x1)              #give correlation matrix of numerical independent variables

vif(x1)              #give variance inflation factors of numerical independent
             #variabels


#outliers and influential observations

halfnorm(rstudent(mod_old))   #half-normal plot of residuals

ga=influence(mod_old)

halfnorm(ga$hat)           #half normal plot of influence

halfnorm(cooks.distance(mod_old))       #half normal plot of Cooks statistics


#logit model excluding possible influential observations:

mod_old1=glm(BAD~.,family=binomial,old_data,subset=c(-556,-1403,-1877,-2508))


cbind(coef(mod_old),coef(mod_old1))     #comparing parameters


#assumptions on new data

x2=model.matrix(mod_new)[,-1]

x2=x2[,c(-4:-9)]

cor(x2)

vif(x2)


halfnorm(rstudent(mod_new))   #residuals

ga=influence(mod_new)

halfnorm(ga$hat)

halfnorm(cooks.distance(mod_new))


mod_new1=glm(BAD~.,family=binomial,new_data,subset=c(-49,-86,-220,-245))


cbind(coef(mod_new),coef(mod_new1))     #comparing parameters
```

```
##############################################################################
##


library(MCMCpack)


information=solve(vcov(mod_old)) #information matrix for logit model on old data

information2=diag(diag(information),17,17)     #diagonal information matrix

#bayesian logistic model on new data with informative prior:

bayes_mod=MCMClogit(BAD~.,data=new_data,burnin=500000,mcmc=10000,tune=0.6,b0=coef
(mod_old),B0=information2,subset=c(-49,-86,-220,-245))


sumb=summary(bayes_mod)

sb=sumb$statistics

sb_coefs=sb[,1]     #coefficients of Bayesian logit model

geweke.diag(bayes_mod)     # check geweke diagnostics

#bayesian logistic model on new data with non-informative prior:

bayes_mod1=MCMClogit(BAD~.,data=new_data,burnin=500000,mcmc=10000,tune=0.6,subset
=c(-49,-86,-220,-245))


sumb1=summary(bayes_mod1)

sb1=sumb1$statistics

sb_coefs1=sb1[,1]

geweke.diag(bayes_mod1)           #check geweke diagnostics

##############################################################################
##

# classification tables for models on test data:

#logistic regression:

y_new=ilogit(predict(mod_new1,test_data))

y_new[y_new>0.48]=1

y_new[y_new<=0.48]=0

new_table=table(test_data$BAD,y_new)
```

```
mod_test=glm(BAD~.,family=binomial,test_data)

x=model.matrix(mod_test)



# Bayesian model with informative prior:

y_bayesian=ilogit(colSums(sb_coefs*t(x)))

y_bayesian[y_bayesian>0.48]=1

y_bayesian[y_bayesian<=0.48]=0

bayesian_table=table(test_data$BAD,y_bayesian)


# Bayesian model with non-informative prior:

y_bayesian1=ilogit(colSums(sb_coefs1*t(x)))

y_bayesian1[y_bayesian1>0.48]=1

y_bayesian1[y_bayesian1<=0.48]=0

bayesian_table1=table(test_data$BAD,y_bayesian1)



##############################################################################
##

#performance on test set with varying sample size for new data:

#Bayesian model with informative prior and logistic regression model:

sample_size=c(50,100,150,200,250,300,350,400,450,500,566)

error_bayes=0

error_new=0

for(i in 1:length(sample_size)){

bayes_mod=MCMClogit(BAD~.,data=new_data[1:sample_size[i],],burnin=500000,mcmc=100
00,tune=0.6,b0=coef(mod_old),B0=information2,subset=c(-49,-86,-220,-245))

sumb=summary(bayes_mod)

sb=sumb$statistics

sb_coefs=sb[,1]

mod_test=glm(BAD~.,family=binomial,test_data)
```

```
x=model.matrix(mod_test)

y_bayesian=ilogit(colSums(sb_coefs*t(x)))

y_bayesian[y_bayesian>0.48]=1

y_bayesian[y_bayesian<=0.48]=0

bayesian_table=table(test_data$BAD,y_bayesian)

error_bayes[i]=(bayesian_table[2]+bayesian_table[3])/sum(bayesian_table)


mod_new=glm(BAD~.,family=binomial,new_data[1:sample_size[i],],subset=c(-49,-86,-
220,-245))

y_new=ilogit(predict(mod_new,test_data))

y_new[y_new>0.48]=1

y_new[y_new<=0.48]=0

new_table=table(test_data$BAD,y_new)

error_new[i]=(new_table[2]+new_table[3])/sum(new_table)

}

#bayesian model with non-informative prior:

sample_size=c(50,100,150,200,250,300,350,400,450,500,566)

error_bayes1=0

for(i in 1:length(sample_size)){

bayes_mod=MCMClogit(BAD~.,data=new_data[1:sample_size[i],],burnin=500000,mcmc=100
00,tune=0.6,subset=c(-49,-86,-220,-245))

sumb=summary(bayes_mod)

sb=sumb$statistics

sb_coefs=sb[,1]

mod_test=glm(BAD~.,family=binomial,test_data)

x=model.matrix(mod_test)

y_bayesian=ilogit(colSums(sb_coefs*t(x)))

y_bayesian[y_bayesian>0.48]=1

y_bayesian[y_bayesian<=0.48]=0

bayesian_table=table(test_data$BAD,y_bayesian)

error_bayes1[i]=(bayesian_table[2]+bayesian_table[3])/sum(bayesian_table)

}
```

```
#plotting the errors of the 3 models with varying sample sizes:

plot(sample_size,error_bayes,type="b",ylim=c(0.12,0.22),xlab="Sample
size",ylab="Error")

lines(sample_size,error_new,type="b",ylim=c(0.12,0.22),pch=3)

lines(sample_size,error_bayes1,type="b",ylim=c(0.12,0.22),pch=4)
```

Appendix B: R output for logistic regression on "old" data

```
Call:

glm(formula = BAD ~ ., family = binomial, data = old_data)


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-2.4270  -0.5444  -0.3238  -0.1186   4.0551


Coefficients:

                Estimate Std. Error z value Pr(>|z|)

(Intercept)   -7.194e+00  5.636e-01 -12.765  < 2e-16 ***

LOAN          -2.367e-05  6.501e-06  -3.642 0.000271 ***

MORTDUE       -3.710e-06  2.284e-06  -1.625 0.104238

VALUE          3.034e-06  1.596e-06   1.902 0.057212 .

REASONHomeImp  2.028e-01  1.349e-01   1.504 0.132632

JOBOffice     -6.819e-01  2.245e-01  -3.038 0.002382 **

JOBOther       1.723e-02  1.786e-01   0.096 0.923139

JOBProfExe     4.760e-02  2.099e-01   0.227 0.820586

JOBSales       4.024e-01  4.245e-01   0.948 0.343111

JOBSelf        4.016e-01  3.799e-01   1.057 0.290496

YOJ           -1.615e-02  9.136e-03  -1.768 0.077093 .

DEROG          7.335e-01  8.062e-02   9.098  < 2e-16 ***
```

```
DELINQ         8.040e-01  6.416e-02  12.530  < 2e-16 ***
CLAGE         -5.223e-03  8.650e-04  -6.038 1.56e-09 ***
NINQ           1.367e-01  3.199e-02   4.272 1.94e-05 ***
CLNO          -2.815e-02  6.787e-03  -4.148 3.36e-05 ***
DEBTINC        1.911e-01  1.378e-02  13.868  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2797.4  on 2758  degrees of freedom
Residual deviance: 1866.7  on 2742  degrees of freedom
AIC: 1900.7

Number of Fisher Scoring iterations: 6
```

Appendix C: R output for logistic regressions on “new” data

C1: Model with all “new” data

```
Call:
glm(formula = BAD ~ ., family = binomial, data = new_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.63880  -0.48115  -0.27481  -0.07361   3.76582

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.615e+00  1.364e+00  -6.317 2.67e-10 ***
LOAN           5.850e-06  1.382e-05   0.423 0.672030
MORTDUE       -6.497e-06  6.965e-06  -0.933 0.350944
```

```
VALUE          1.618e-06  5.845e-06   0.277 0.781878
REASONHomeImp  1.087e-01  3.223e-01   0.337 0.736028
JOBOffice     -9.802e-01  5.819e-01  -1.684 0.092110 .
JOBOther       1.622e-01  4.551e-01   0.357 0.721458
JOBProfExe     1.056e-01  5.287e-01   0.200 0.841722
JOBSales       3.329e+00  9.417e-01   3.535 0.000408 ***
JOBSelf       -1.441e-01  9.040e-01  -0.159 0.873381
YOJ           -2.675e-02  2.150e-02  -1.244 0.213402
DEROG          6.563e-01  2.100e-01   3.125 0.001779 **
DELINQ         1.157e+00  1.675e-01   6.904 5.05e-12 ***
CLAGE         -6.654e-03  2.082e-03  -3.196 0.001394 **
NINQ           2.056e-01  6.585e-02   3.122 0.001798 **
CLNO          -4.076e-02  1.629e-02  -2.501 0.012374 *
DEBTINC        2.327e-01  3.331e-02   6.985 2.86e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 568.67  on 565  degrees of freedom
Residual deviance: 341.18  on 549  degrees of freedom
AIC: 375.18

Number of Fisher Scoring iterations: 6
```

C2: Model with influential observations removed

```
Call:
glm(formula = BAD ~ ., family = binomial, data = new_data, subset = c(-49,
    -86, -220, -245))
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.70684  -0.45126  -0.24970  -0.05463   2.83375

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.973e+00  1.513e+00  -6.594 4.28e-11 ***
LOAN           6.060e-06  1.499e-05   0.404 0.685952
MORTDUE       -2.016e-06  8.223e-06  -0.245 0.806292
VALUE         -1.263e-06  6.933e-06  -0.182 0.855411
REASONHomeImp  8.325e-02  3.376e-01   0.247 0.805237
JOBOffice     -9.882e-01  6.088e-01  -1.623 0.104515
JOBOther       2.696e-01  4.759e-01   0.566 0.571069
JOBProfExe     1.538e-02  5.593e-01   0.027 0.978067
JOBSales       4.162e+00  1.012e+00   4.111 3.94e-05 ***
JOBSelf        4.595e-02  9.349e-01   0.049 0.960803
YOJ           -2.699e-02  2.240e-02  -1.205 0.228258
DEROG          1.075e+00  3.210e-01   3.348 0.000815 ***
DELINQ         1.214e+00  1.789e-01   6.786 1.16e-11 ***
CLAGE         -7.775e-03  2.192e-03  -3.547 0.000390 ***
NINQ           1.900e-01  6.876e-02   2.762 0.005739 **
CLNO          -4.136e-02  1.738e-02  -2.379 0.017338 *
DEBTINC        2.694e-01  3.716e-02   7.249 4.20e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 564.11  on 561  degrees of freedom
Residual deviance: 314.87  on 545  degrees of freedom
AIC: 348.87
```

```
Number of Fisher Scoring iterations: 6
```

Appendix D: Bayesian logistic regression models on "new" data

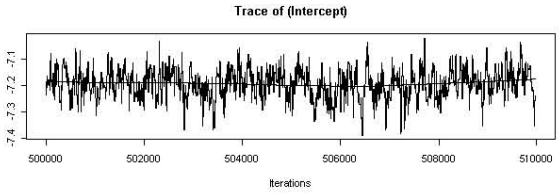
D1: R output for Bayesian logistic regression model on "new" data with informative prior

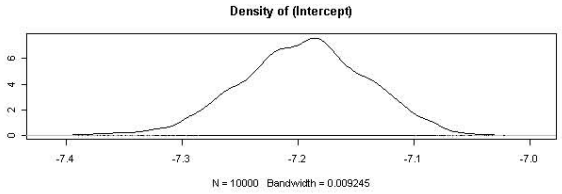
```
Iterations = 500001:510000

Thinning interval = 1

Number of chains = 1

Sample size per chain = 10000


1. Empirical mean and standard deviation for each variable,

   plus standard error of the mean:


                     Mean        SD  Naive SE Time-series SE

(Intercept)    -7.196e+00 5.515e-02 5.515e-04      4.191e-03

LOAN           -2.234e-05 2.897e-06 2.897e-08      2.562e-07

MORTDUE        -3.858e-06 6.771e-07 6.771e-09      5.126e-08

VALUE           3.033e-06 4.891e-07 4.891e-09      3.637e-08

REASONHomeImp   1.706e-01 1.008e-01 1.008e-03      6.946e-03

JOBOffice      -7.141e-01 1.481e-01 1.481e-03      1.125e-02

JOBOther        3.359e-02 8.635e-02 8.635e-04      6.542e-03

JOBProfExe      1.505e-02 1.267e-01 1.267e-03      1.008e-02

JOBSales        8.206e-01 3.475e-01 3.475e-03      2.600e-02

JOBSelf         3.111e-01 3.002e-01 3.002e-03      2.493e-02

YOJ            -1.583e-02 5.211e-03 5.211e-05      3.773e-04

DEROG           7.360e-01 7.200e-02 7.200e-04      5.737e-03

DELINQ          8.348e-01 5.013e-02 5.013e-04      3.216e-03

CLAGE          -5.216e-03 3.206e-04 3.206e-06      2.264e-05

NINQ            1.456e-01 2.350e-02 2.350e-04      1.915e-03

CLNO           -2.799e-02 2.573e-03 2.573e-05      2.014e-04
```

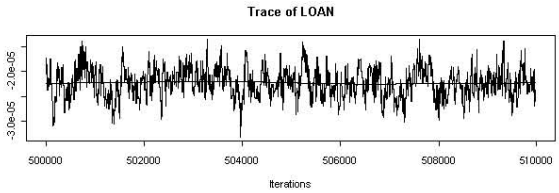
```
DEBTINC        1.915e-01 1.381e-03 1.381e-05      9.573e-05
```

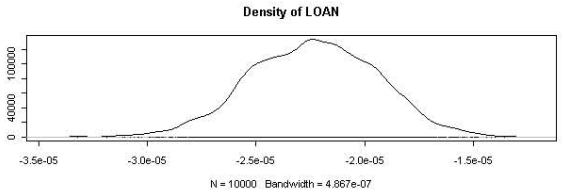
2. Quantiles for each variable:

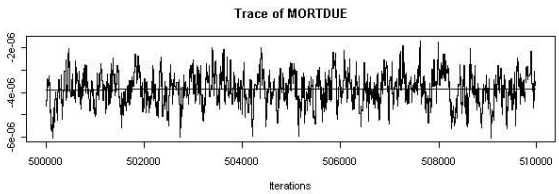
```
                    2.5%        25%        50%        75%      97.5%
(Intercept)   -7.305e+00 -7.231e+00 -7.193e+00 -7.158e+00 -7.091e+00
LOAN          -2.808e-05 -2.439e-05 -2.233e-05 -2.028e-05 -1.680e-05
MORTDUE       -5.270e-06 -4.298e-06 -3.842e-06 -3.417e-06 -2.566e-06
VALUE          2.095e-06  2.723e-06  3.021e-06  3.338e-06  4.027e-06
REASONHomeImp -2.591e-02  1.015e-01  1.749e-01  2.394e-01  3.671e-01
JOBOffice     -1.003e+00 -8.096e-01 -7.220e-01 -6.085e-01 -4.216e-01
JOBOther      -1.383e-01 -2.490e-02  3.577e-02  9.502e-02  1.980e-01
JOBProfExe    -2.194e-01 -6.997e-02  1.381e-02  9.808e-02  2.766e-01
JOBSales       1.123e-01  5.975e-01  8.296e-01  1.052e+00  1.482e+00
JOBSelf       -3.238e-01  1.214e-01  3.123e-01  5.160e-01  8.704e-01
YOJ           -2.560e-02 -1.941e-02 -1.589e-02 -1.243e-02 -4.653e-03
DEROG          5.930e-01  6.886e-01  7.375e-01  7.833e-01  8.820e-01
DELINQ         7.357e-01  8.012e-01  8.348e-01  8.653e-01  9.377e-01
CLAGE         -5.859e-03 -5.423e-03 -5.212e-03 -4.998e-03 -4.623e-03
NINQ           1.004e-01  1.285e-01  1.449e-01  1.629e-01  1.906e-01
CLNO          -3.320e-02 -2.965e-02 -2.800e-02 -2.619e-02 -2.310e-02
DEBTINC        1.889e-01  1.906e-01  1.915e-01  1.924e-01  1.944e-0
```

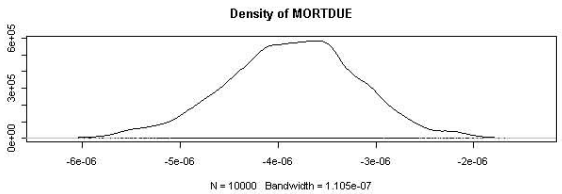

Trace of (Intercept)
Iterations

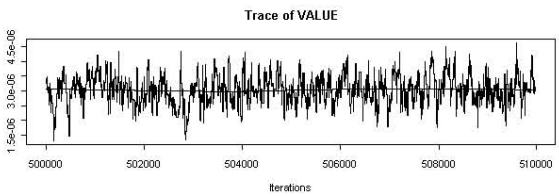

Density of (Intercept)
N = 10000 Bandwidth = 0.009245

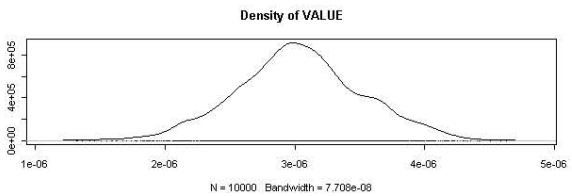

Trace of LOAN
Iterations

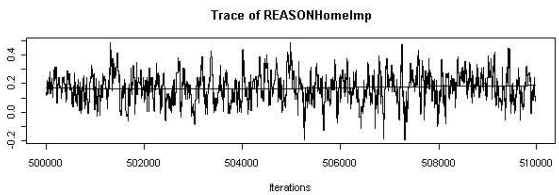

Density of LOAN
N = 10000 Bandwidth = 4.867e-07

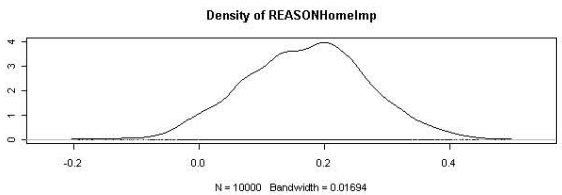

Trace of MORTDUE
Iterations

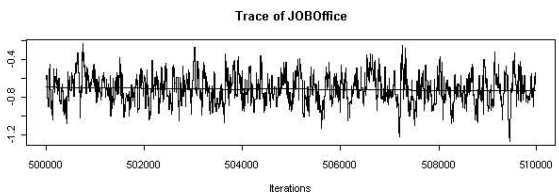

Density of MORTDUE
N = 10000 Bandwidth = 1.105e-07

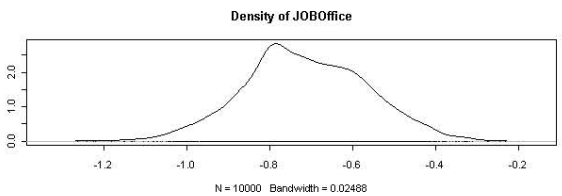

Trace of VALUE
Iterations

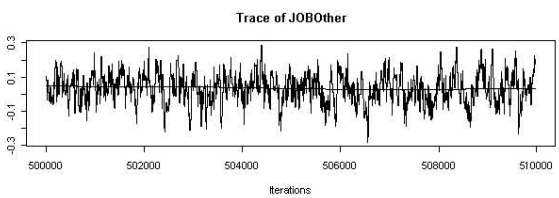

Density of VALUE
N = 10000 Bandwidth = 7.708e-08

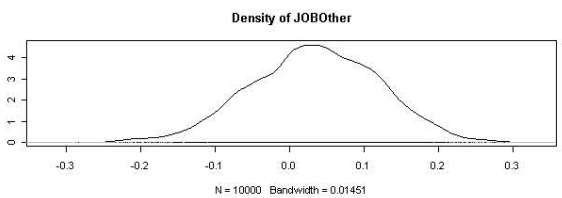

Trace of REASONHomeImp
Iterations

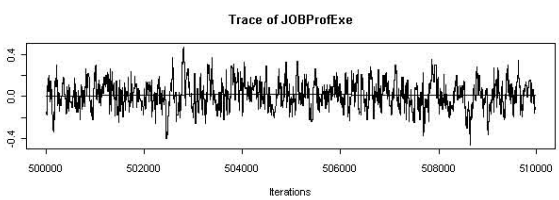

Density of REASONHomeImp
N = 10000 Bandwidth = 0.01694

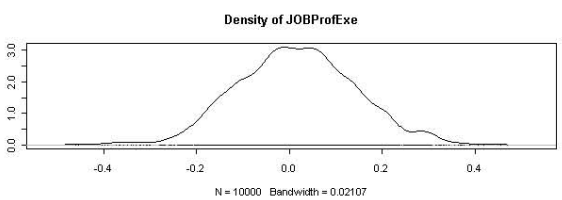

Trace of JOBOffice
Iterations

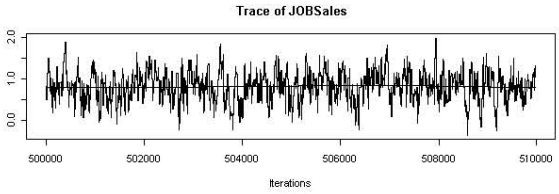

Density of JOBOffice
N = 10000 Bandwidth = 0.02488

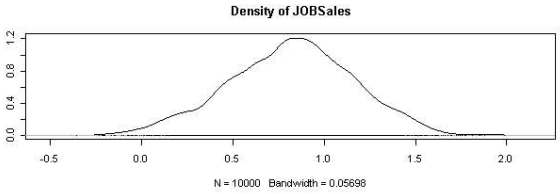

Trace of JOBOther
Iterations

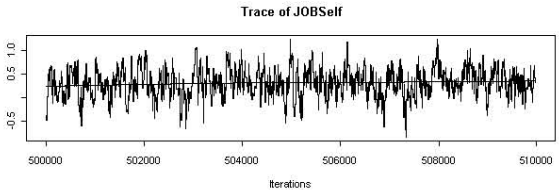

Density of JOBOther
N = 10000 Bandwidth = 0.01451

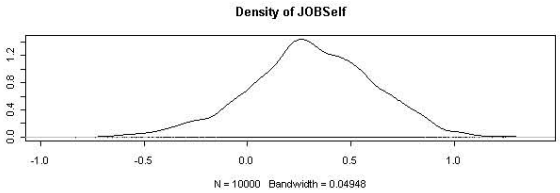

Trace of JOBProfExe
Iterations

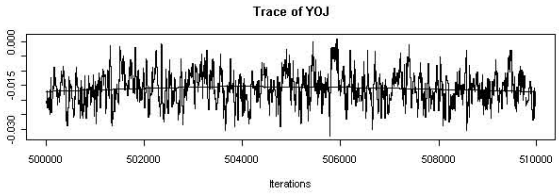

Density of JOBProfExe
N = 10000 Bandwidth = 0.02107

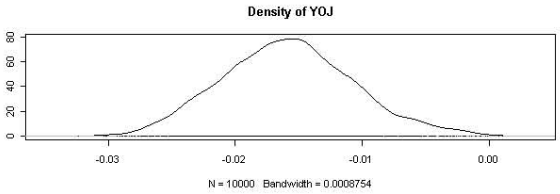

Trace of JOBSales
Iterations

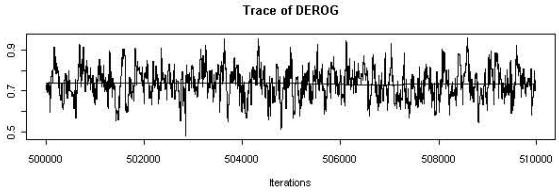

Density of JOBSales
N = 10000 Bandwidth = 0.05698

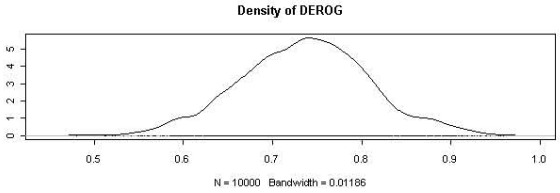

Trace of JOBSelf
Iterations

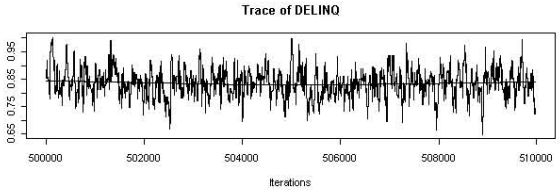

Density of JOBSelf
N = 10000 Bandwidth = 0.04948

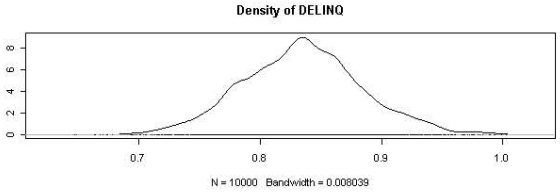

Trace of YOJ
Iterations

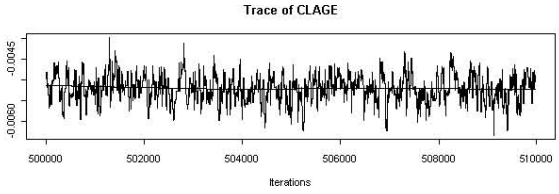

Density of YOJ
N = 10000 Bandwidth = 0.0008754

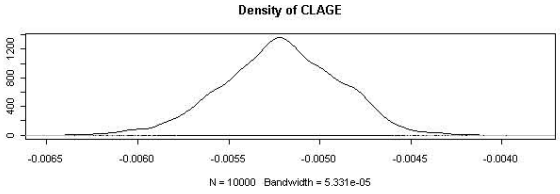

Trace of DEROG
Iterations

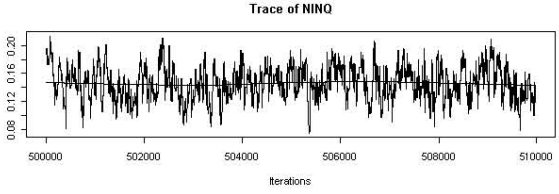

Density of DEROG
N = 10000 Bandwidth = 0.01186

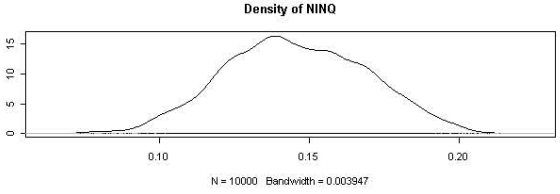

Trace of DELINQ
Iterations

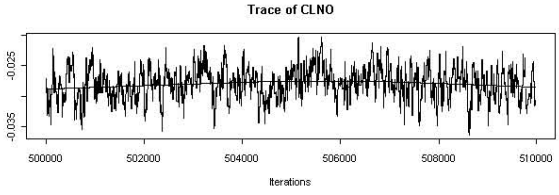

Density of DELINQ
N = 10000 Bandwidth = 0.008039

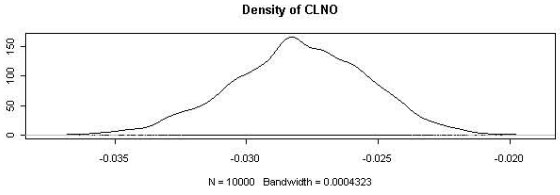

Trace of CLAGE
Iterations

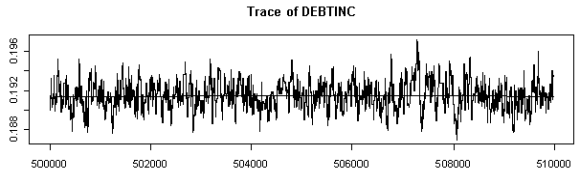

Density of CLAGE
N = 10000 Bandwidth = 5.331e-05

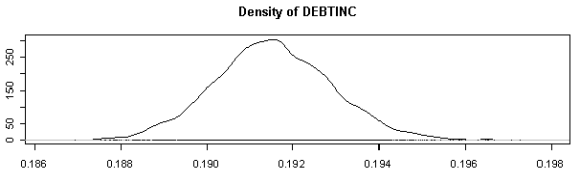

Trace of NINQ
Iterations


Density of NINQ
N = 10000 Bandwidth = 0.003947


Trace of CLNO
Iterations


Density of CLNO
N = 10000 Bandwidth = 0.0004323





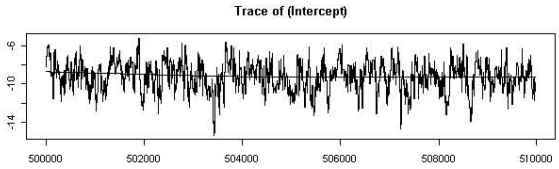
D2: R output for Bayesian logistic regression model on "new" data with non-informative prior

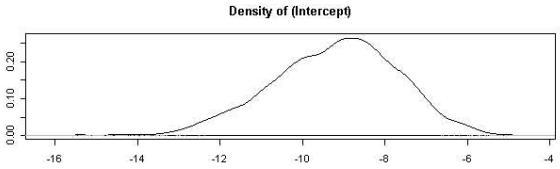
```
Iterations = 500001:510000

Thinning interval = 1

Number of chains = 1

Sample size per chain = 10000


1. Empirical mean and standard deviation for each variable,

   plus standard error of the mean:


                    Mean        SD  Naive SE Time-series SE

(Intercept)    -9.212e+00 1.538e+00 1.538e-02      1.258e-01

LOAN            5.248e-06 1.398e-05 1.398e-07      1.135e-06

MORTDUE        -7.716e-06 6.984e-06 6.984e-08      5.136e-07

VALUE           2.173e-06 5.727e-06 5.727e-08      4.358e-07

REASONHomeImp   9.077e-02 3.562e-01 3.562e-03      2.747e-02

JOBOffice      -9.888e-01 5.947e-01 5.947e-03      4.724e-02

JOBOther        1.646e-01 4.717e-01 4.717e-03      3.383e-02

JOBProfExe      1.176e-01 5.555e-01 5.555e-03      3.766e-02

JOBSales        3.568e+00 9.672e-01 9.672e-03      7.448e-02

JOBSelf        -2.271e-01 9.678e-01 9.678e-03      8.455e-02

YOJ            -2.852e-02 2.233e-02 2.233e-04      1.652e-03

DEROG           7.688e-01 2.209e-01 2.209e-03      1.639e-02

DELINQ          1.236e+00 1.742e-01 1.742e-03      1.204e-02
```

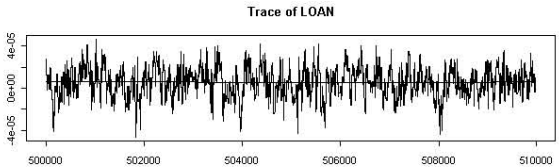
CLAGE	-7.204e-03	2.083e-03	2.083e-05	1.461e-04
NINQ	2.123e-01	7.219e-02	7.219e-04	5.723e-03
CLNO	-4.485e-02	1.744e-02	1.744e-04	1.458e-03
DEBTINC	2.515e-01	3.642e-02	3.642e-04	3.043e-03

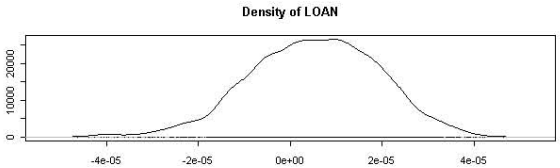
2. Quantiles for each variable:

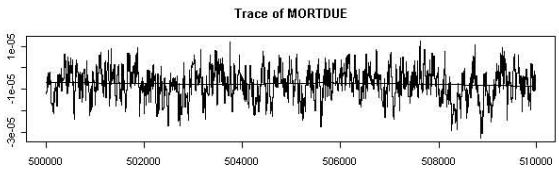
	2.5%	25%	50%	75%	97.5%
(Intercept)	-1.241e+01	-1.022e+01	-9.093e+00	-8.134e+00	-6.470e+00
LOAN	-2.320e-05	-4.472e-06	5.762e-06	1.545e-05	3.118e-05
MORTDUE	-2.196e-05	-1.203e-05	-7.241e-06	-2.907e-06	5.288e-06
VALUE	-9.056e-06	-1.559e-06	2.280e-06	5.789e-06	1.395e-05
REASONHomeImp	-5.866e-01	-1.532e-01	9.478e-02	3.365e-01	7.699e-01
JOBOffice	-2.101e+00	-1.392e+00	-9.925e-01	-5.842e-01	1.343e-01
JOBOther	-7.109e-01	-1.476e-01	1.516e-01	4.850e-01	1.116e+00
JOBProfExe	-9.444e-01	-2.559e-01	1.032e-01	4.617e-01	1.279e+00
JOBSales	1.643e+00	2.937e+00	3.529e+00	4.213e+00	5.488e+00
JOBSelf	-2.182e+00	-8.519e-01	-1.890e-01	4.236e-01	1.624e+00
YOJ	-7.150e-02	-4.308e-02	-2.810e-02	-1.333e-02	1.415e-02
DEROG	3.742e-01	6.119e-01	7.583e-01	9.061e-01	1.271e+00
DELINQ	8.960e-01	1.116e+00	1.239e+00	1.349e+00	1.581e+00
CLAGE	-1.112e-02	-8.553e-03	-7.233e-03	-5.860e-03	-3.071e-03
NINQ	6.882e-02	1.633e-01	2.130e-01	2.609e-01	3.505e-01
CLNO	-7.773e-02	-5.703e-02	-4.547e-02	-3.242e-02	-1.056e-02
DEBTINC	1.894e-01	2.248e-01	2.482e-01	2.748e-01	3.295e-01

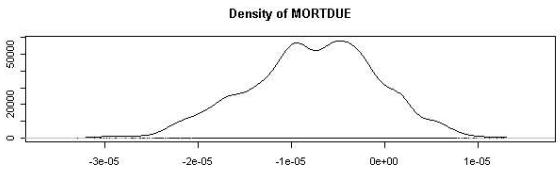

Trace of (Intercept)
Iterations

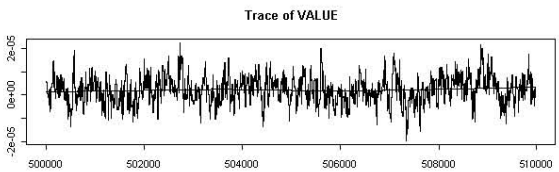

Density of (Intercept)
N = 10000 Bandwidth = 0.2584

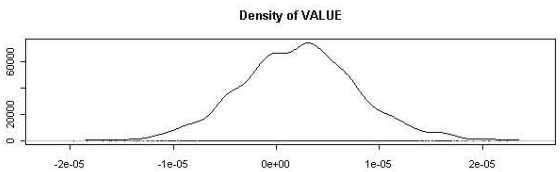

Trace of LOAN
Iterations

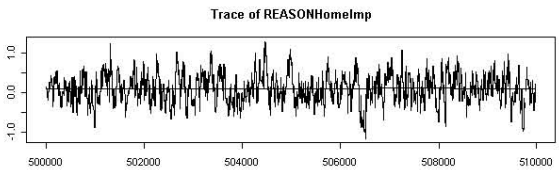

Density of LOAN
N = 10000 Bandwidth = 2.349e-06

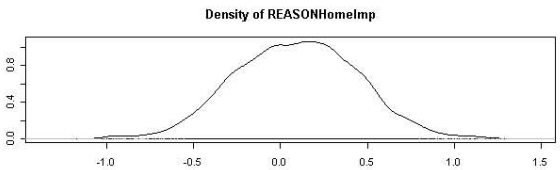

Trace of MORTDUE
Iterations

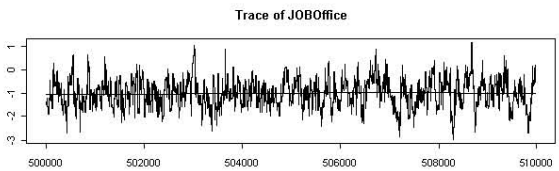

Density of MORTDUE
N = 10000 Bandwidth = 1.144e-06

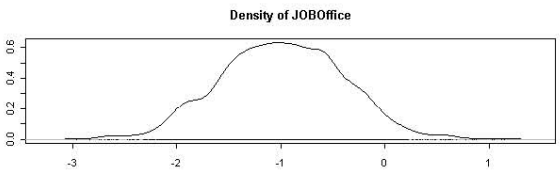

Trace of VALUE
Iterations

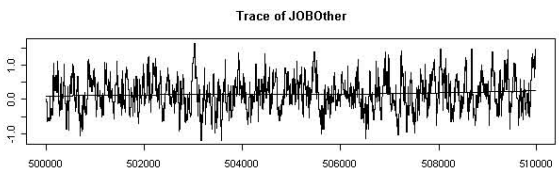

Density of VALUE
N = 10000 Bandwidth = 9.213e-07

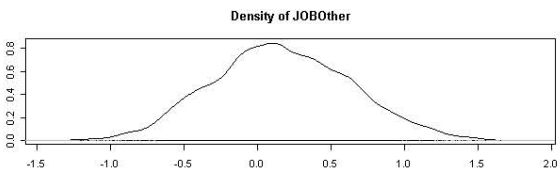

Trace of REASONHomeImp
Iterations

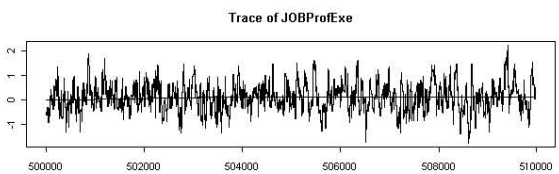

Density of REASONHomeImp
N = 10000 Bandwidth = 0.05984

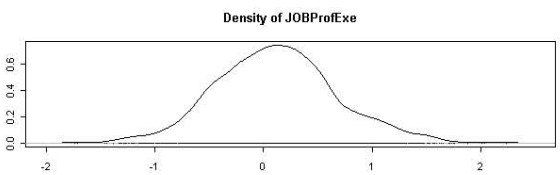

Trace of JOBOffice
Iterations

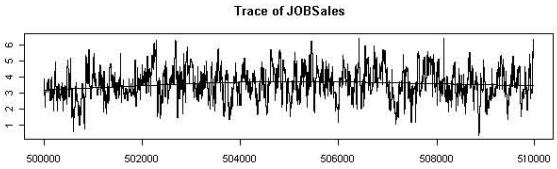

Density of JOBOffice
N = 10000 Bandwidth = 0.09991

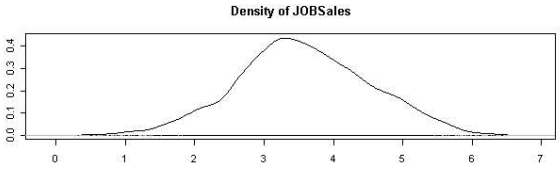

Trace of JOBOther
Iterations

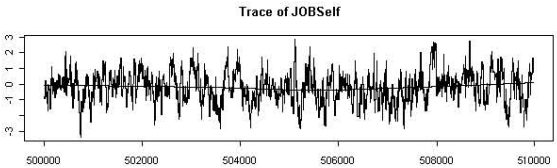

Density of JOBOther
N = 10000 Bandwidth = 0.07925

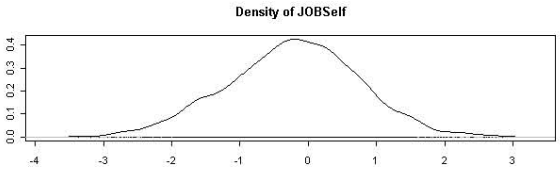

Trace of JOBProfExe
Iterations

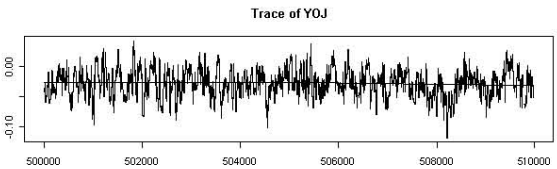

Density of JOBProfExe
N = 10000 Bandwidth = 0.08997

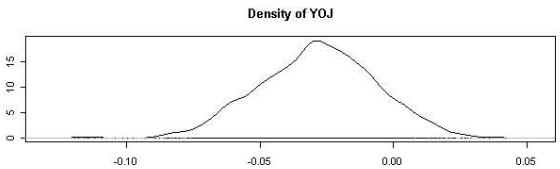

Trace of JOBSales
Iterations

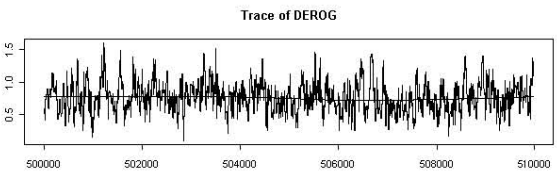

Density of JOBSales
N = 10000 Bandwidth = 0.16

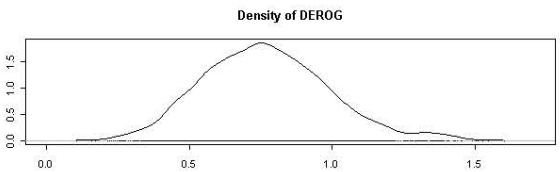

Trace of JOBSelf
Iterations

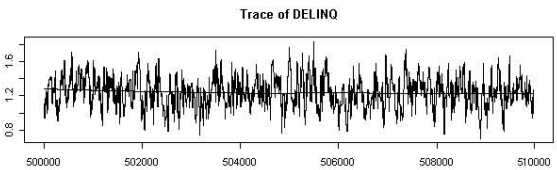

Density of JOBSelf
N = 10000 Bandwidth = 0.1599

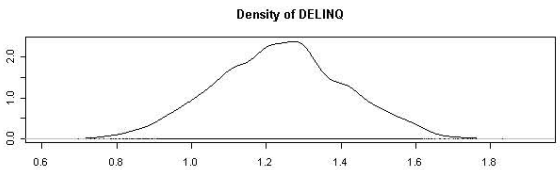

Trace of YOJ
Iterations

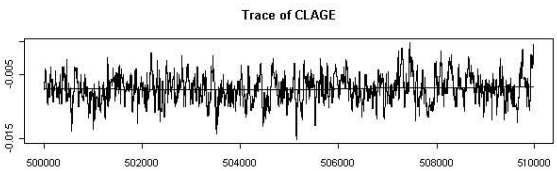

Density of YOJ
N = 10000 Bandwidth = 0.00373

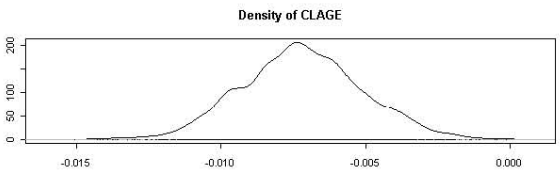

Trace of DEROG
Iterations

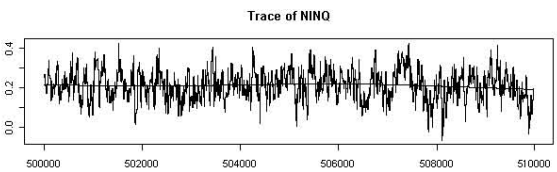

Density of DEROG
N = 10000 Bandwidth = 0.03689

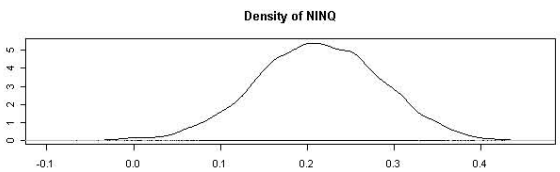

Trace of DELINQ
Iterations

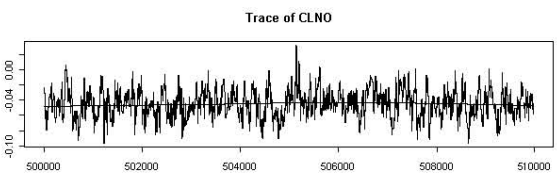

Density of DELINQ
N = 10000 Bandwidth = 0.02927

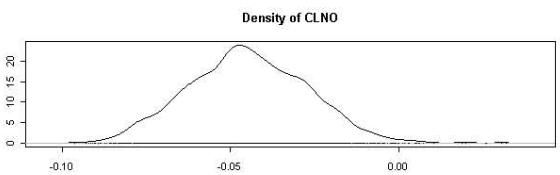

Trace of CLAGE
Iterations

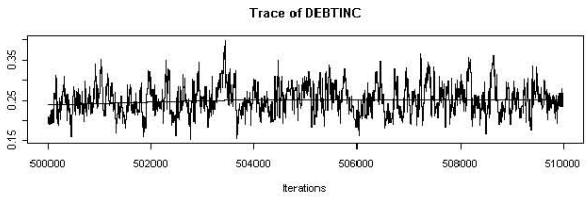

Density of CLAGE
N = 10000 Bandwidth = 0.0003376

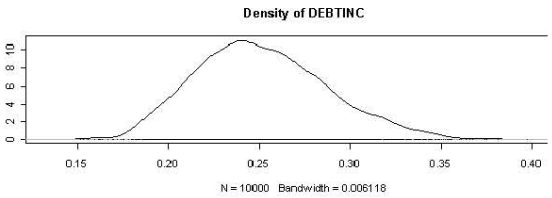

Trace of NINQ
Iterations


Density of NINQ
N = 10000 Bandwidth = 0.01213


Trace of CLNO
Iterations


Density of CLNO
N = 10000 Bandwidth = 0.002931


Trace of DEBTINC
0.15
0.25
0.35
500000
502000
504000
506000
508000
510000
Iterations


Density of DEBTINC
0
2
4
6
8
10
0.15
0.20
0.25
0.30
0.35
0.40
N = 10000 Bandwidth = 0.006118