A Bayesian Analysis of a Random Effects Small Business Loan Credit Scoring Model

Patrick J. Farrell
School of Mathematics and Statistics
Carleton University, 1125 Colonel By Drive
Ottawa, Ontario, CANADA
pfarrell@math.carleton.ca

Brenda MacGibbon
Département de mathématiques
Université du Québec à Montréal
C.P. 8888, Succursale Centre-Ville
Montréal, Québec, CANADA
macgibbon.brenda@uqam.ca

Thomas J. Tomberlin
Sprott School of Business
Carleton University, 1125 Colonel By Drive
Ottawa, Ontario, CANADA
Jerry_Tomberlin@carleton.ca

Dale Doreen
Department of Decision Sciences and
Management Information Systems
John Molson School of Business
Concordia University
1455 de Maisonneuve Blvd. West
Montréal, Québec, CANADA
doreen@jmsb.concordia.ca

Abstract

One of the most important aspects of credit scoring is constructing a model that has low misclassification rates and is also flexible enough to allow for random variation. It is also well known that, when there are a large number of highly correlated variables as is typical in studies involving questionnaire data, a method must be found to reduce the number of variables to those that have high predictive power. Here we propose a Bayesian multivariate logistic regression model with both fixed and random effects for small business loan credit scoring and a variable reduction method using Bayes factors. The method is illustrated on an interesting data set based on questionnaires sent to loan officers in Canadian banks and venture capital companies.

Keywords: Bayes Factors, Credit Scoring, MCMC, Variable Selection.

1. Introduction

The health and growth of the small business sector is of pivotal importance to both the banking communities and the economies of developing and developed countries. Thus sufficient credit for this sector must be made available and tools

must be developed for loan officers, since the evaluation of a small business requires a greater degree of judgement (Levin and Travis (1987)). The analysis of financial statements is the starting point in almost any evaluation of a loan application. However, with small businesses, standard financial statements may not provide enough information to make an informed decision about the company. The consideration of qualitative factors such as the personal, entrepreneurial, and managerial attributes of the people become more important in the loan appraisal (Cooper 1975 and Hays 1977). This necessitates separate credit decision guidelines for small business loans. The quantification of these factors is known as credit scoring. Thomas et al. (2002) provide a good overview of credit scoring, while Blöchlinger and Leippold (2006) demonstrate the economic benefit of good credit scoring.

Credit scores are quantitative measures that try to capture the risk associated with granting the loan. There is increased research interest in credit scores for different types of loans. Studies are often organized and analyzed in order to assist in the decision making process for loan officers. They are often based on questionnaire data as well as financial statements. For example, Doreen and Farhoomand (1983) conduct such a study and use discriminant analysis techniques to analyze the data. However, one of the assumptions of discriminant analysis is that the predictive variables have an approximate multivariate normal distribution within categories of the outcome or dependent variable. While this may be the case with continuous variables, it is certainly not necessarily true for ordered categorical variables. The multivariate logistic function does not share the same problem and provides an excellent alternative to linear discriminant procedures. Fournier et al. (1994) successfully use such a function in an analysis of a small business loan data set in order to determine the bias in apparent misclassification rates using bootstrap methods.

Although it is important to have a credit scoring model with high predictive value, it is also important to account for random variation in the data. As the outcome variable is often a binary one representing whether or not the loan is granted, hierarchical logistic regression can be used with both fixed and random effects. Many applications of the hierarchical logistic regression model originally proposed by Wong and Mason (1985) exist, although we concentrate here on those related to banking and credit scoring. Leonard (1993) uses empirical Bayes methods on a mixed effects model in order to evaluate the commercial loan process. Hand and Henley (1997) provide an overview of statistical methodology useful for consumer credit scoring. Avery et al. (2000) use logistic regression to study future loan performance and conclude that credit scoring improves the efficiency for the review process relative to solely depending on credit bureau scores. Kocenda and Vojtek (2009) compare parametric (logistic regression) and nonparametric CART (classification and regression trees) approaches for credit scoring models, and find them to be very similar. They also conclude that socio-demographic variables should not be excluded from the model. Khudnitskaya (2010) introduces a multilevel model to predict the probability of loan defaults in retail banking and compares her procedure to standard logistic regression using ROC curves (Hanley and McNeill 1982). Brentnall et al. (2010) successfully use

a multinomial random-effects model in a linear prediction model for the amount withdrawn at cash machines.

One of the major drawbacks in the analysis of questionnaire data for constructing credit scoring models is the large number of variables (that is, the number of questions) compared to the number of individuals who respond. This necessitates some type of variable selection regardless of whether Bayesian or non-Bayesian techniques are used.

In the situation where there are a large number of correlated predictors, there exists many interesting Bayesian approaches to building parsimonious linear prediction models (some of which are being adapted to generalized linear models such as logistic regression). These include the use of Gibbs sampling techniques, the choice of relevant mixture priors, stochastic search variable selection algorithms, Markov chain Monte Carlo and/or model mixing. These methodologies are accurately summarized for our purposes here in the work of Carlin and Chib (1995), Kass and Rafferty (1995) and George and McCulloch (1997). However, Tüchler (2008) recently proposes a very interesting Markov chain Monte Carlo algorithm using a stochastic search variable approach to select explanatory variables and to determine the random effects covariance matrix in a mixed effects logistic regression model.

Here we use a data set originally collected by Decheverry and Doreen (1985), which considers qualitative and quantitative criteria in devising decision models of the small business loan decision. Using proportional sampling, loan officers in banks and venture capital companies across Canada are sent a questionnaire (see the Appendix) asking them to evaluate, using a categorical interval scale, the importance of various qualitative and quantitative factors in their acceptance or rejection of recent small business loan applications. The questionnaire consists of 27 questions that were intended to evaluate four particular economic and administrative sectors; nine of these questions (M1 through M9) focused on the qualities of the manager of the business, seven (P1 through P7) on the earning potential of the enterprise, five (S1 through S5) on the level of the security/risk associated with the loan, and six (E1 through E6) on the economic and commercial environment of the business. Each question is evaluated on a five point scale from 1 (very poor / high risk) to 5 (excellent / low risk). We use these data here to illustrate the utility of the variable selection method for the hierarchical Bayes techniques proposed here for a logistic regression model with random and fixed effects. There are several goals in our analysis. Important ones are to determine which variables should be included in the model, as well as to estimate the individual parameters that determine this choice and to properly account for the random variation in the data. Since variable reduction is essential to building a good credit scoring model, another important goal is to introduce the use of Bayes factors to accomplish such reduction for the hierarchical Bayes logistic model with fixed and random effects.

The paper is organized as follows. Section 2 presents the hierarchical model with fixed and random effects that is used to analyze this type of data. It also

describes how to obtain hierarchical Bayes estimates for such a model using adaptive rejection Metropolis sampling within the Gibbs sampler of Gilks et al. (1995). Section 3 discusses our method of variable selection. In Sections 4 and 5, we describe the data collected by Decheverry and Doreen (1985), and analyze these data using both the standard logistic regression model and the more complex hierarchical Bayes logistic regression model with fixed and random effects. We follow this with a conclusion and discussion of our approach in Section 6.

2. The Hierarchical Bayes Logistic Model with Mixed Effects and Related *Parameter Estimation*

We restrict our attention here to the analysis of data leading to a dichotomous choice. More precisely, let Y_j = 1 if the *j*-th loan application is approved, and let Y_j = 0 if it is not. The random variable Y_j follows a Bernoulli distribution with parameter π_j, reflecting the probability that the *j*-th loan application is approved. To study the effects of various fixed effect covariates on the decision to grant a loan or not, one possibility is to use a standard logistic regression model (Model 1):

$$Y_j \sim \textit{Bernoulli}\ (\pi_j),$$

$$\textit{logit}\ (\pi_j) = \textit{ln}\left(\frac{\pi_j}{1-\pi_j}\right) = x_j'\beta, \qquad (1)$$

where x_j is a vector, augmented by the constant one, of covariates associated with the *j*-th loan application (note that x_j' in (1) is the transpose of x_j) while β is an associated parameter vector which contains a constant term β_0. All 27 variables in the data set to be analyzed are possible candidates for the model, depending upon how strongly each of them affects the loan assessment process. Maximum likelihood estimation of the parameters in (1) is straightforward and can be accomplished using any statistical software package such as MINITAB or SAS. We choose to use MINITAB, Version 16 here.

However, this model may not describe the data well if there are covariates related to the loan assessment process other than those in x_j that influence whether or not a loan is approved. If this is the case, a random effects logistic regression model may be more appropriate. In order to consider such a situation here, we must decide if there exists an apt random effect that can be derived from our questionnaire. Avery et al. (2000) argue that there are two different types of scores – those that reflect information found in various financial records and those that take into account factors other than financial history. It is this second type of score, which is often obtained from questionnaires which is of primary interest here. This paper focuses on the analysis of such data. However, let us examine our questionnaire carefully. There are four categories of variables: management, earning potential, security (financial) and environment. It is mainly the first, second and fourth categories that are qualitative, while the third does

contain hard financial data. The advantage of the third category is that it contains the type of data that is collected by almost any loan officer and if a random variable is constructed from questions in this category with several different levels, then these levels can be envisaged as a sample from a larger population of levels and can be considered as a reflection of random variation in loan applicants. This reasoning leads us to create a random effect variable consisting of five levels by combining the values of two out of the five security variables in the data set (we shall describe this in more detail in Section 5). We also assume that the remaining three security variables are not observed, thus leaving only the management, earning potential, and environmental variables (22 in total) as possible candidates for the covariate vector. We feel that these 22 variables correspond more closely to Avery et al.'s (2000) definition of qualitative attributes excluding financial history.

For the random effects logistic regression model, we require a notation that is slightly more complicated than for the model in equation (1). Specifically, we let Y_{ij} = 1 if the *j*-th loan with security level *i* is approved, and let Y_{ij} = 0 otherwise. As above, Y_{ij} follows a Bernoulli distribution, this time with parameter π_{ij}, reflecting the probability that the *j*-th loan with security level *i* is granted. The random effects logistic regression model (Model 2) is given by

$$\begin{aligned} Y_{ij} &\sim \textit{Bernoulli}\ (\pi_{ij}), \\ \textit{logit}\ (\pi_{ij}) &= \textit{ln}\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x_{ij}'\beta + \delta_i, \\ \delta_i &\sim \textit{Normal}\ (0,\ \sigma^2). \end{aligned} \tag{2}$$

where x_{ij} is a vector, augmented by the constant one, of covariates associated with the *j*-th loan application with security level *i*. The quantity δ_i is a random effect associated with the *i*-th security level. We assume that these effects are normally distributed with unknown variance σ^2.

One possible approach to the estimation of the parameters in (2) is an empirical Bayes procedure using the EM algorithm of Dempster et al. (1977) and incorporating the approximation proposed by Laird (1978) to avoid the intractable numerical integration as is done in Farrell et al. (1997). Alternatively, a hierarchical Bayes approach may also be used, and we choose to use such a procedure here. This approach requires the specification of prior distributions for σ^2 and the parameters in the β vector. We consider here a diffuse version of an inverse gamma distribution for the random effects variance σ^2 and a flat priors for the parameters in β.

The computational method often used is the Gibbs sampler, originally proposed by Geman and Geman (1984). An excellent discussion of this method can be found in Gelfand and Smith (1990). In order to obtain the hierarchical Bayes parameter estimates, Markov chain Monte Carlo (MCMC) is used here as in Farrell (2000). Although software such as WinBUGS is now readily available for

such Bayesian computations and the model proposed here could be implemented in it, we include a short description of the method we use for simulating the posterior distributions of the model parameter estimates.
Hierarchical Bayes estimation procedures for the parameters in the model given by (2) require knowledge about the posterior distributions of these parameters. However, it is only possible to know these distributions up to a constant of proportionality; specifically, the posterior distribution for any given parameter is proportional to the product of all terms in the model that contain it. Therefore, for Model (2), if *Y* and δ are vectors containing Y_{ij} and δ_i respectively, while *X* is a matrix with rows x'_{ij}, then

$$f(\beta_0 \mid Y,\beta_1,\ldots,\beta_m,\delta,\sigma^2,X) \;\alpha\; \prod_{ij}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}},$$

$$f(\beta_u \mid Y,\beta_0,\beta_1,\ldots,\beta_{u-1},\beta_{u+1},\ldots,\beta_m,\delta,\sigma^2,X) \;\alpha\; \prod_{ij}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}},$$

$$f(\delta_i \mid Y,\beta,\delta_1,\ldots,\delta_{i-1},\delta_{i+1},\ldots,\delta_n,\sigma^2,X) \;\alpha\; \prod_{ij}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}\;\exp(-\frac{1}{2}\sum_i\frac{\delta_i^2}{\sigma^2}),$$

$$f(\sigma^2 \mid Y,\beta,\delta,X) \;\alpha\; \frac{1}{\tau^{n+2}}\exp(-\frac{1}{2}\sum_i\frac{\delta_i^2}{\sigma^2}),$$

where *u* refers to the *u*-th covariate, *m* is the number of covariates, and *n* is the number of security levels.

Under Gibbs sampling, an initial set of values are assumed as the estimates for β, δ, and σ^2, say $\hat{\beta}_{\{0\}}$, $\hat{\delta}_{\{0\}}$, and $\hat{\sigma}^2_{\{0\}}$. An updated estimate for β_0, say $\hat{\beta}_{0\{1\}}$, is obtained by sampling from the full conditional distribution $f(\beta_0 \mid Y,\hat{\beta}_{1\{0\}},\ldots,\hat{\beta}_{m\{0\}},\hat{\delta}_{\{0\}},\hat{\sigma}^2_{\{0\}},X)$. Sampling from the full conditional distribution $f(\beta_1 \mid Y,\hat{\beta}_{0\{1\}},\hat{\beta}_{2\{0\}},\ldots,\hat{\beta}_{m\{0\}},\hat{\delta}_{\{0\}},\hat{\sigma}^2_{\{0\}},X)$ based on $\hat{\beta}_{0\{1\}}$ yields the revised estimate $\hat{\beta}_{1\{1\}}$ for β_1. The completion of a first iteration is realized once the revised estimates $\hat{\beta}_{\{1\}}$, $\hat{\delta}_{\{1\}}$, and $\hat{\sigma}^2_{\{1\}}$ are obtained. This procedure of sampling using the most up-to-date revised estimates continues until the estimates of each parameter are deemed to have stabilized from one iteration to the next. See Geman and Geman (1984) and Gelfand and Smith (1990) for a discussion on Gibbs sampling, and Gelman and Rubin (1992) for methods of convergence.

A different full conditional distribution is sampled each time a new estimate is obtained; regardless of which parameter is being estimated. Since many iterations are usually needed to ensure that estimates for each parameter have stabilized, efficient methods for constructing full conditional distributions and sampling from them are required. For log-concave distributions, this can be done using adaptive rejection sampling (See Gilks and Wild 1992). For applications where the full conditional distributions are not log-concave, when the number of cases associated with certain random effects is very small, or when the random effects distribution is highly skewed, Gilks et al. (1995) propose appending a Hasting-Metropolis algorithm step to the adaptive rejection scheme, and suggest

using the resulting adaptive rejection Metropolis sampling scheme within the Gibbs sampling algorithm. We follow this approach here.

3. Variable Selection Procedures

The credit scoring data set for small business loans to be analyzed here contains many covariates. As discussed in the previous section, there are twenty-seven possible candidates available for inclusion in Model (1), and twenty-two for Model (2). In this section, we describe the procedures that we employ for variable selection in each of the two models. For both models, we utilize forward selection procedures in which variables are added one at a time until none of the remaining covariates are deemed to improve the model.

One variable selection method that can be used for Model (1) is based on a Wald test statistic (See Wald 1943). For any given variable included in the model, this statistic is simply the ratio of the estimate of the associated model parameter in the β vector divided by its estimated asymptotic standard error. Under a null hypothesis that this particular parameter in the β vector equals zero, this test statistic is distributed approximately as a standard normal for large sample sizes. Thus, it is straightforward to determine a P-value for testing this null hypothesis. For fitting Model (1) using maximum likelihood estimation, all statistical software packages provide the estimate for β, the estimated asymptotic covariance matrix of the estimator for β, along with the P-values for conducting a Wald test on each covariate in the model.

For Model (1), a forward selection procedure based on the Wald statistic for determining which variables to include in the model can proceed as follows. Initially, we consider fitting every possible model based on (1) that would include one and only one covariate. The variable that yields the smallest P-value for the associated Wald statistic is the first to be added to the model, provided that this P-value is smaller than some pre-specified level of significance, which is usually chosen to be 0.10, 0.05, or 0.01. Here we choose 0.05. The procedure then continues by fitting every possible two-variable model that would consist of the first variable that was added and one of the remaining variables. The variable of those remaining that yields the smallest P-value for the associated Wald statistic is the next to be included in the model, again provided that this P-value is smaller than the pre-specified level of significance. We continue to add variables to the model in this fashion until we encounter the situation where the smallest P-value exceeds the specified significance level. When this occurs, the variable linked to this P-value is not added to the model, and the selection procedure is deemed complete.

Variable selection for Model (2) fit using a hierarchical Bayes approach can proceed in the same forward selection fashion as that employed for Model (1). However, rather than using P-values based on Wald statistics that allow us to

add variables one at a time, we can make use of Bayes factors to determine whether or not to add a variable, and if so, which one. An advantage of Bayes factors over frequentist hypothesis tests which can only reject the null hypothesis is that Bayes factors can assess the strength of the evidence in favour of the null (See Kass and Raftery 1995). Another advantage of Bayes factors is that they naturally guard against overfitting by including a penalty for including too much model structure. We now describe the computation of a Bayes factor for choosing between two models that contain different sets of covariates.

As above, suppose that data on Y_{ij}, representing whether or not the *j*-th loan with security level *i* is approved, are summarized in the vector *Y*. We assume that these data arise under one of two models, say a model with covariate vectors $x_{ij(1)}$ or another with a different set of covariate vectors, $x_{ij(2)}$, according to a probability density $f(Y \mid X_{(1)})$ or $f(Y \mid X_{(2)})$, respectively. Here $X_{(k)}$ is a matrix with rows $x'_{ij(k)}$, for *k* = 1, 2. Using Bayes Theorem, the Bayes factor can be shown to be the ratio of these two densities, namely $B_{12} = f(Y \mid X_{(1)}) / f(Y \mid X_{(2)})$. Note that these two densities are obtained by integrating over the parameter space rather than the usual maximizing of the numerator and denominator which is used in frequentist calculations. A value of B_{12} greater than one suggests a model with covariate vectors $x_{ij(1)}$ is more strongly supported by the data than one with vectors $x_{ij(2)}$. Jeffreys (1961) gives a series of recommendations for the interpretation of B_{12}, which are based on half units on a logarithmic (base 10) scale. These recommendations are summarized in Table 1.

Table 1: Guidelines for interpreting the Bayes factor, B_{12}, according to Jeffreys (1961)

Value of B_{12}	Evidence Against Model With Vectors $x_{ij(2)}$
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
Greater than 100	Decisive

4. The Data

We consider a data set arising from a study conducted by Decheverry and Doreen (1985) in the John Molson School of Business at Concordia University in Montreal, Quebec, Canada. Specifically, a questionnaire concerning small business loan requests is circulated to loan officers in Canadian banks and venture capital companies across the country. The questionnaire consists of 27 questions that were intended to evaluate four particular economic and administrative sectors; nine of these questions (M1 through M9) focus on the qualities of the manager of the business, seven (P1 through P7) on the earning potential of the enterprise, five (S1 through S5) on the level of the security/risk associated with the loan, and six (E1 through E6) on the economic and

commercial environment of the business. The response to each question is evaluated on a five point scale from 1 (very poor / high risk) to 5 (excellent / low risk). The questionnaire also collected information on whether the loan was approved or not. Of the 151 questionnaires completed, 110 loans are granted, yielding an approval rate of 73%. Finally, of note is the fact that 34 questionnaires has at least one missing value regarding the twenty seven questions used to evaluate the four economic and administrative sectors. The majority of these questionnaires, 22, are only missing the answer to one question. Of the remaining twelve questionnaires with missing values, there are four with two missing, three each with three and four missing, and one each with five and six missing. Prior to analyzing the data, the missing values for a particular question are replaced with the modal value obtained from all of the responses to that question.

5. The Analysis

In the analysis that we perform on the data described in the previous section, we consider the decision to grant a loan or not to be the Bernoulli response, assigning a value of one if the loan is approved, and zero if not. Initially, we use a forward selection procedure based on the Wald test statistic to choose variables for the standard logistic regression model in (1). The significance level was set at 0.05. The statistical software package MINITAB, Version 16 is used to fit the standard logistic regression models. Five variables entered, in the following order:

S5: Applicant rating regarding security based on past information (P-value = 0.000)
P4: Degree of accuracy in cash flow projection (P-value = 0.000)
E2: Degree of competition who supply a similar product/service (P-value = 0.006)
E1: Degree of development of the market for the product/service (P-value = 0.022)
E4: Degree of definition of the distribution system for the product/service (P-value = 0.042)

At this stage, the variable that could most improve the model with these five variables already included was S2, reflecting the proportion of the business financed by the owner/manager. However, the associated P-value, albeit close to the significance level, is slightly larger at 0.056. Table 2 reports the fits of two standard logistic regression models. Both of these models contain S5, P4, E2, E1, and E4; however, one also includes S2, while the other does not.

Table 2: Parameter estimates for the standard logistic regression model after variable selection. Standard errors are given in brackets after the parameter estimate

	Without S2	With S2
Variable	Estimate (Std Err)	Estimate (Std Err)
Constant	-15.114 (2.863)	-15.600 (3.007)
S5	1.6202 (0.4471)	1.3882 (0.4584)

P4	1.8242 (0.4555)	1.7496 (0.4878)
E2	0.6597 (0.3268)	0.5016 (0.3365)
E1	0.8596 (0.4188)	0.9004 (0.4323)
E4	0.6367 (0.3133)	0.6480 (0.3299)
S2	---	0.6140 (0.3207)

Next, we consider variable selection in the context of a random effects logistic regression model as given in (2) using Bayes factors. In order to do so, we create a random effect variable consisting of five categories describing the level of security associated with a loan. First, for each loan assessment observation in the data set, we sum the values associated with S2 and S5 (both of which were measured on a 5 point scale), which yield a value somewhere between 2 and 10, inclusive, for each observation. We then group the loan assessment observations into five categories distinguished by whether the sum of S2 and S5 for a particular observation was at most 3, 4, 5, 6, or at least 7. These five categories, labeled SRE1 through SRE5, are subsequently used to define security level random effects for the model given in (2), since the higher the sum of S2 and S5, the greater the perceived security of the loan. Finally, when it comes to deciding which variables to include as possible candidates for entry into the model as fixed effects, we opt to assume that the remaining three security variables are unobserved; hence the creation of the random effect for security level. Thus, we only consider management, earning potential, and environmental variables (22 in total) as candidates for the covariate vector.

In fitting the random effects logistic regression model given in (2), the procedure employed by Gilks et al (1995) is used. Specifically, each time that the model is fit with a particular set of covariates, the Gibbs sampler is run for 15,000 iterations twice, each with a different set of starting values for the parameter estimates. The method of Gelman and Rubin (1992) is used to assess convergence of the Gibbs sampler. To ensure proper convergence, only the last 3000 iterations of each of the two runs are used to construct posterior distributions. Specifically, the results over these two sets of 3000 iterations are combined in order to approximate these distributions.

In order to determine the variables to include in the fixed effects covariate vector in the mixed effects logistic regression model, we employ a selection procedure based on Bayes factors. Initially, we perform pairwise comparisons of a model containing only the random effects for security level to every possible model containing one fixed effect covariate and the same random effect terms. Of all these comparisons, the model with P4, the degree of accuracy in cash flow projection, produces the highest Bayes factor, 16.7, when compared to the model only containing random effects. This variable is therefore included as the first in the fixed effect covariate vector. We then continue using a similar approach to determine if additional variables can be added to the fixed effect covariate vector, and if so, which one(s). Including P4, three variables are deemed worthy, entering the model in the following order:

P4: Degree of accuracy in cash flow projection (B_{12} = 16.7)

E1: Degree of development of the market for the product/service (B_{12} = 6.39)

E4: Degree of definition of the distribution system for the product/service (B_{12} = 4.03)

At this stage, the variable that can most improve the random effects model with these three variables already included as fixed effects is M5, reflecting the experience of the manager in a different industry. However, the associated Bayes factor, albeit close to 3.2, is smaller at 2.84. Table 3 reports the fits of two random effects logistic regression models. Both of these models contain the random effects for security level, P4, E1, and E4; however, one also includes M5, while the other does not. Regarding these estimates, two points are worthy of note. First, the random effect for security level seems to be important in both models, as the ratio of the estimate to the standard error is quite large (greater than 2) for SRE5. Second, in the model including M5, the estimate of the regression coefficient for this variable is negative. This would suggest that the greater the experience of the manager in a different industry, the less likely the loan will be granted.

Table 3: Parameter estimates for the random effects logistic regression model after variable selection. Standard errors are given in brackets after the parameter estimate

	Without M5	With M5
Variable	Estimate (Std Err)	Estimate (Std Err)
Constant	-10.0440 (2.6048)	-9.9492 (2.8523)
SRE1	-2.2015 (1.3568)	-2.5456 (1.5226)
SRE2	-1.2291 (1.0697)	-1.9099 (1.0721)
SRE3	-0.1129 (1.0358)	-0.0108 (1.0622)
SRE4	1.5620 (1.0365)	2.0485 (1.1600)
SRE5	1.9852 (0.9290)	2.4148 (1.0395)
P4	1.6838 (0.5302)	1.7844 (0.5551)
E1	0.9625 (0.4505)	1.1290 (0.4891)
E4	0.7483 (0.3650)	0.9204 (0.4149)
M5	---	-0.6669 (0.3937)

Note that for each of the four model fits summarized in Tables 2 and 3, it is possible to use the parameter estimates to compute, for each of the 151 loan applications comprising the data set, an estimate of the probability that the loan will be approved according to the model. Suppose that we decide to use a rule to grant a loan if the model predicts a probability of 0.5 or greater, of the loan being granted. We can then compare how each of the four models will perform in terms of forecasting the decision that was actually made on each loan application, which is also available from the data set. For each of the four models in Tables 2 and 3, Table 4 presents misclassification rates that are based on the rule above. For example, consider the standard logistic regression model without S2. Using the 0.5 probability rule above, this model incorrectly predicts granting 6 of the 41 loans (14.63%) that are not granted, and not granting 3 of the 110

loans (5.45%) that are granted. The model therefore predicts incorrectly on 9 of the 151 loan applications, for an overall error rate of 7.95%. Given the error rates summarized in Table 4, the random effects model that includes M5 is the best for classification. It should be noted that numerically the overall misclassification rates for both of the hierarchical Bayes logistic regression models with fixed and random effects are lower than those obtained using either of the standard logistic regression models here. Note also that the inclusion of S2 in the standard logistic regression model improves the effectiveness of the model to correctly identify when a loan is granted, while not affecting its ability to detect when a loan is not approved. This is despite the fact that the decision criterion for including S2 in the model suggests that it can be excluded. A similar conclusion can be drawn when M5 is added to the random effects model.

Table 4: Misclassification rates for the standard logistic regression and random effects models

Model	Loan Not Approved	Loan Approved	Overall Error Rate
Logistic without S2	0.1463	0.0545	0.0795
Logistic with S2	0.1463	0.0273	0.0596
Random effects without M5	0.0976	0.0364	0.0530
Random effects with M5	0.0976	0.0273	0.0464

6. Conclusion and Discussion

Our premise here is that a Bayesian approach to credit scoring for small business loans would be advantageous. We test this by comparing a hierarchical Bayes multivariate logistic regression model with both fixed and random effects for credit scoring and a variable reduction method using Bayes factors with the usual logistic regression model using forward selection for variable reduction on the same data. The results of this analysis indicate to us that the Bayesian methodology is worth pursing in future applications. There are three directions we feel that the future research in credit scoring for small business loans could profitably take.

First it would be important to design a new questionnaire taking into account the more recent research into credit scoring models. The inclusion of more socio-demographic variables and geographic variables seems to us essential given the work of Avery et al. (2000), DeYoung et al. (2008), Kocenda and Vojtek (2009) and Khudnitskaya (2010). For example, Avery et al. (2000) show the importance of not omitting local geographic economic variables for lending in the real estate market. DeYoung et al. (2008) also study the effects on loan performance of borrower-lender distance and whether credit scoring was used. They determine that on average, both these variables, appropriately coded, are separately associated with higher default probabilities, but that the use of hard financial information reduced the distance default-increasing effects. Kocenda and Vojtek (2009) introduce nonparametric CART (classification and regression trees) into

credit scoring. They also conclude that socio-demographic variables should not be excluded from the model.

The recent research of Khudnitskaya (2010) on the prediction of the probability of loan defaults in retail banking should be incorporated into research on credit scoring models for small business loans. She introduces a multilevel model for this prediction problem and compares her procedure with standard logistic regression using ROC curves (See Hanley and McNeill 1982). This comparison will be important in future research on modeling the small business loan process. She also accounts for unobserved characteristics of the degree to which the client is worthy of credit by introducing random effects. The second level, which she calls the microenvironment, contained information about the geographic area where the client lived along with other socio-demographic and socioeconomic variables.

Secondly, more recent Bayesian approaches to variable selection could be incorporated into the analysis of the questionnaires associated with small business credit scoring. The stochastic search variable selection algorithms, the specification of a hierarchical Bays mixture prior and the general area of model mixing for building parsimonious linear prediction models discussed in George and McCullogh (1997) should be adapted to logistic regression and exploited in model building for credit scoring. The work of Tüchler (2008), who develops a very interesting Markov chain Monte Carlo algorithm using a stochastic search variable approach to select explanatory variables in a mixed effects logistic regression model, should be helpful for this.

Lastly, Fournier et al. (1994) use the bootstrap successfully to determine the bias in apparent misclassification rates for the standard logistic regression model. The Bayesian method for variable selection described here uses Bayes factors for variable selection and is more computer intensive. An analogue for the bootstrap in this Bayesian setting would be very useful.

Appendix: "Small Business Loan Checklist" Questionnaire

Please rate the loan applicant according to the following factors, on a scale of 1 to 5 (1 = very poor, 5 = excellent).

I. MANAGEMENT

- M1. Owner/Manager's administrative abilities
- M2. Owner/Manager's technical and operational abilities
- M3. Owner/Manager's decision-making abilities
- M4. Owner/Manager's experience in same industry
- M5. Owner/Manager's experience in different industry
- M6. Owner/Manager's educational attainment
- M7. Owner/Manager's ambition and drive

M8. Owner/Manager's communication skills

M9. Owner/Manager's innovativeness

M10. In relation to this loan, what is the most important managerial factor of those stated above?

II. EARNING POTENTIALS

P1. Sales forecasts are realistic

P2. Projected income figures are accurate and realistic

P3. Payment plan is realistic

P4. Cash flow projection is accurate and realistic

P5. Forecasted earnings can meet the loan payments

P6. Future expansion can be financed through the projected earnings

P7. Expected growth rate is realistic

P8. In relation to this loan, what is the most important financial criteria of those stated above?

III. SECURITY

S1. Sufficient security is offered to protect the loan

S2. The proportion of business financed by the owner/manager

S3. The security offered is liquid

S4. The quality of the security is reflected by the deterioration of its value

S5. According to available information regarding security (previous loans, etc.), how do you rate the applicant?

S6. In relation to this loan, what is the most important security factors of those stated above?

IV. ENVIRONMENT

E1. The market for this product/service is: developed/growing (less risky = 5) or undeveloped/stagnant (more risky = 1)

E2. The existing competition who supply similar product service are: plentiful (more risky = 1) or scarce (less risky = 5)

E3. Suppliers for owner/manager's product/service are: closely situated and plentiful (less risky = 5) or scarce and distant (more risky = 1)

E4. Owner/manager has distribution system for his product/service clearly defined: YES (less risky = 5) or NO (more risky = 1)

E5. Nature of product/service is marketable: YES/proven (less risky = 5) or NO/undetermined (more risky = 1)

E6. General economic trend at time is conducive to product/services marketability: NO (more risky = 1) or YES (less risky = 5)

E7. In relation to this loan, what is the most important environment factor of those stated above?

V. SUMMARY

1. Was the loan granted? YES NO
2. In relation to this loan, which of the factors had most influence on the final decision, please rank (least important = 4, most important = 1):

   management skills ____
   earnings potential ____
   security of loan ____
   environment ____

3. If the loan was granted, give the amount in dollars: $
4. During your career as a credit manager, what percentage of small business loans, in your opinion, have successfully been collected?
5. What percentage of small business loan applications, in your opinion, are accepted?
6. If you wish to receive a copy of the research findings, please provide name and address.

Acknowledgements

The authors are grateful to two referees for helpful comments. The authors would also like to thank D. Morin, G. Brighten and M. Decheverry who participated earlier in different aspects of the research project. The research of the first and second authors of this paper was supported by NSERC of Canada.

References

1. Avery, R.B., Bostic, R.W., Calem, P.S., and Canner G.B. (2000) “Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files”, *Real Estate Economics*, 28, 523-547.
2. Blöchlinger, A., and Leippold, M. (2006), “Economic Benefit of Powerful Credit Scoring”, *Journal of Banking & Finance*, 30, 851-873.
3. Brentnall, A.R., Crowder, M.J., and Hand, D.J. (2010), “Predicting the Amount Individuals Withdraw at Cash Machines Using a Random Effects Multinomial Model”, *Statistical Modelling*, 10, 197-214.
4. Carlin, B.P., and Chib, S. (1995), “Bayesian Model Choice via Markov Chain Monte Carlo Methods”, *Journal of the Royal Statistical Society B*, 57, 473-484.

5. Cooper, P.J. (1975) “Four P’s for Lending”, *Journal of Commercial Bank Lending*, 46-50.

6. Decheverry, M., and Doreen, D. (1985), “Survey and Analysis of Criteria Used in Assessing Small Business Credit”, MBA Business Research Project, Concordia University.

7. Dempster, A.P., Laird N.M., and Rubin, D.B. (1977), “Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm”, *Journal of the Royal Statistical Society B*, 39, 1-38.

8. DeYoung, R., Glennon, D., and Nigro, P. (2008), “Borrower–Lender Distance, Credit Scoring, and Loan Performance: Evidence from Informational-Opaque Small Business Borrowers”, *Journal of Financial Intermediation*, 17, 113-143.

9. Doreen, D., and Farhoomand, F. (1983), “A Decision Model for Small Business Loans”, *Journal of Small Business*, 1, 18-28.

10. Farrell, P.J. (2000), “Bayesian Inference for Small Area Proportions”. *Sankhya B*, 62, 402-416.

11. Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1997), “Empirical Bayes Estimators of Small Area Proportions in Multistage Designs”, *Statistica Sinica*, 7, 1065-1083.

12. Fournier, B., Larribé, F., MacGibbon, B., Morin, D., and Doreen, D. (1994), “L’évaluation du biais dans le taux d’erreur apparent d’une régression logistique suivant la méthode du bootstrap”, *Actes du colloque sur les méthodes et domaines d’application de la statistique*, Bureau de la Statistique du Québec 1994, 219-222.

13. 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.

14. Gelman, A., and Rubin, D.B. (1992), “Inference from Iterative Simulations Using Multiple Sequences”, *Statistica Sinica*, 7, 457-511.

15. Geman, S., and Geman, D. (1984), “Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

16. George E., and McCulloch, R. (1997), “Approaches for Bayesian Variable Selection”, *Statistica Sinica*, 7, 339-373.

17. Gilks, W.R., Best, N.G., and Tan, K.K. (1995), “Adaptive Rejection Metropolis Sampling Within Gibbs Sampling”, *Applied Statistics*, 44, 455-472.

18. Gilks, W.R., and Wild, P. (1992), “Adaptive Rejection Sampling for Gibbs Sampling”, *Journal of the Royal Statistical Society C*, 41, 337-348.

19. Hand, D., and Henley, W. (1997), "Statistical Classification Methods in Consumer Credit Scoring: A Review", *Journal of the Royal Statistical Society A*, 160, 523-541.

20. Hanley, J.A., and McNeil, B.J. (1982), "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve", *Radiology*, 143, 29-36.

21. Hays, R.S. (1977), *Business Loans: A Guide to Money Sources and How to Approach Them Successfully*, John Wiley and Sons: New York, 1977.

22. Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford U.K.: Oxford University Press, 1961.

23. Kass, R.E. and Raftery, A.E. (1995), "Bayes Factors", *Journal of the American Statistical Association*, 90, 773-795.

24. Khudnitskaya, A.S. (2010), "Improved Credit Scoring with Multilevel Statistical Modelling", PhD Thesis, Technischen Universität Dortmund, 2010.

25. Kocenda, E., and Vojtek, M. (2009), "Default Predictors and Credit Scoring Models for Retail Banking", *CESIFO Working Paper No. 2862, Category 12: Empirical and Theoretical Methods*.

26. Laird, N.M. (1978), "Empirical Bayes Methods for Two-Way Contingency Tables", *Biometrika*, 65, 581-590.

27. Leonard, K.J. (1993), "Empirical Bayes Analysis of the Commercial Loan Evaluation Process", *Statistics & Probability Letters*, 18, 289-296.

28. Levin, R.I., Travis V.R. (1987), "Small Company Finance: What the Books Don't Say", *Harvard Business Review*, 30-32.

29. Thomas, L.C., Edelmann, D.B., and Crook, J.N. (2002), "Credit Scoring and its Applications", SIAM, Philadelphia, 2002.

30. Tüchler, R.(2008), "Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling", *Journal of Computational and Graphical Statistics*, 17, 76-94.

31. Wald, A. (1943), "Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large", *Tranactions of the American Mathematical Society*, 54, 426-482.

32. Wong, G.Y., and Mason, W.M. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis", *Journal of the American Statistical Association*, 80, 513-524.