Monte Carlo simulation study of regression models used to estimate the credit banking risk in home equity loans

D. Morales[1], A. Pérez-Martín[2] & M. Vaca[2]
[1]*Operations Research Center, Miguel Hernández University of Elche, Spain*
[2]*Department of Economics and Finance, Miguel Hernández University of Elche, Spain*

Abstract

The banking structure of today is quite damaged. This happened because the industry was not able to foresee the different risks that surrounded it. Of the group of risks associated with the business of banking activity, the risk of credit in many occasions accounts for 60%. The risk of credit arises when there exists the possibility of suffering a loss due to the breach of the other party to assume the payment or payments. The default originates a loss for the entity that climbs not only to the none recovered amount, but also to the expenses incurred in the process. The uncertain nature of the risk does mean that this risk is measured through the unexpected loss, which coincides statistically with the standard deviation. This is why statistical methods are needed to enable the prediction of bank credit risk (default and non-payment) in home equity loans through estimates based on statistical models (also called techniques of 'credit scoring'), to improve the currently available methods.
Keywords: credit scoring, credit risk, home equity loans, linear mixed models, Monte Carlo.

1 Introduction

To assess credit risk, there are a variety of methodologies available, from the personalized study of an experienced risk analyst to different statistical and econometrics methods of credit scoring. Credit scoring is essentially a way to identify different groups in a population. The first proposal to solve this problem

WIT Transactions on Information and Communication Technologies, Vol 45, © 2013 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA130131

was introduced in [1] using discriminant analysis as a multivariate statistical technique. Durand [2] was the first in admitting that the same statistical techniques can be used to optimize the differentiation between good and bad loans. When a credit counselor assesses the risk they are assigning a score to the risk of a loan application from their past experience, i.e. they are applying a score. Credit scoring models or methods are algorithms and a evaluation of credit risk automatically occurs. They have a single dimension.

From 1970 credit scoring models used were based on statistical techniques (in particular, discriminant analysis), but this is then generalized from 1990. The best statistical resources were developed with the use of technology and there is a growing need on the part of financial institutions to make more effective and efficient funding and a better risk assessment. Today, credit scoring models are based on mathematics, econometric techniques and artificial intelligence. During the late 20th century and early 21st, there was economic growth. Consumer credit was increased spectacularly. Evaluation of credit risk by statistical methods become outstanding [3]. This author make a study of judgmental versus scoring methods in score-cards application. Historically, these are discriminant analysis and linear regression logistic regression, probit analysis, nonparametric smoothing methods, mathematical programming, Markov chain models, recursive partitioning, expert systems, genetic algorithms, neural networks and conditional independence models.

Empirical studies by various authors present alternative approaches for comparing different techniques. The study by [4] compared various techniques and found that decision trees outperform logistic regressions, as these yield better results than discriminant analysis. The studies of [5], show the inferiority of the parametric models against the non-parametric ones in analyzing their predictive quality. The analysis of [6] concluded that neural networks development is significantly better than the discriminant analysis, while [7] obtained reversed results. Yatchew [8] carried out an in-depth study which analyzes the advantages and disadvantages of the use of non-parametric regression techniques. Comparing parametric and non-parametric methods the study points out the non-existence of an optimal method for all portfolios. Studies of [9] also agree with the designated authors. In the analysis of several types of default [10] include the hypothesis of instability. Some references in which the statistical methodology of discriminant analysis is also used in the credit scoring problem are [11] and [12].

Therefore, the scientific literature has still not solved the problem efficiently using a method through estimates based on statistical models (also called credit scoring techniques) in order to improve those currently available for home equity loans. In this paper we propose procedures for fitting linear regression models. We start with a fixed effects model. On the other hand, we consider a mixed (fixed and random effects) model. To evaluate the appropriateness of these procedures, Monte Carlo simulation experiments are preformed that allow the comparison of certain properties between them. Searle *et al.* [13] provided a detailed description of linear mixed models.

Our main goal is to discern which kind of linear regression model (mixed or fixed effects) is more appropriate to fit datasets from home equity loans. For this reason, we purpose two models in section 2. In this section we develop maximum likelihood (ML) and residual maximum likelihood (REML) estimation methods. We describe an exhaustive simulation study of the models in section 3. Finally, in section 4 conclusions are offered.

2 The models

Let us consider the following linear mixed model with a random effect as the main model. The random effect has I levels ($i = 1, \ldots, I$), and each level i has n_i units. The model is

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + u_i + w_{ij}^{-1/2}e_{ij}, \quad i = 1, \ldots, I,\ j = 1, \ldots, n_i, \tag{1}$$

where y_{ij} is the value of the target variable in the population unit j at the level i, $\boldsymbol{x}_{ij}$ is a row vector of fixed effects containing the values of p auxiliary and linearly independent variables, w_{ij} is a known heteroscedasticity weight and $\boldsymbol{\beta}$ is a column vector of regression parameters. Further, the random effects u_i and the errors e_{ij} are assumed to be mutually independent with distributions $u_i \sim N(0, \sigma_1^2)$ and $e_{ij} \sim N(0, \sigma_0^2)$, respectively. The model (1) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{W}^{-1/2}\boldsymbol{e}, \tag{2}$$

where $\boldsymbol{u} = \boldsymbol{u}_{1,I\times 1} \sim N_I(\boldsymbol{0}, \sigma_1^2\boldsymbol{I}_I)$, $\boldsymbol{e} = \boldsymbol{e}_{n\times 1} \sim N_n(\boldsymbol{0}, \sigma_0^2\boldsymbol{I}_n)$ are independent, $\boldsymbol{y} = \boldsymbol{y}_{n\times 1}$, $\boldsymbol{X} = \underset{1\le i\le I}{\text{col}}(\boldsymbol{X}_i)$ with $\text{rank}(\boldsymbol{X}) = p$, $\boldsymbol{X}_i = \underset{1\le j\le n_i}{\text{col}}(\boldsymbol{x}_{ij})$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p\times 1}$, $\boldsymbol{Z} = \underset{1\le i\le I}{\text{diag}}(\boldsymbol{1}_{n_i})$, $n = \sum_{i=1}^{I} n_i$, $\boldsymbol{I}_a$ is the identity matrix of order a, $\boldsymbol{1}_a$ is the column vector of dimension a whose elements are all equal to 1, $\boldsymbol{W} = \underset{1\le i\le I}{\text{diag}}(\boldsymbol{W}_i)$, $\boldsymbol{W}_i = \underset{1\le j\le n_i}{\text{diag}}(w_{ij})$ with $w_{ij} > 0$ known. Under model (2), it holds that $\boldsymbol{V} = var(\boldsymbol{y}) = \boldsymbol{Z}var(\boldsymbol{u})\boldsymbol{Z}^t + \sigma_0^2\boldsymbol{W}^{-1} = \underset{1\le i\le I}{\text{diag}}(\boldsymbol{V}_i)$, where $\boldsymbol{V}_i = \sigma_1^2\boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^t + \sigma_0^2\boldsymbol{W}_i^{-1}$. Then it follows that $y \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$.

If σ_0^2 and σ_1^2 are known, the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y}$, but usually it isn't the reality.

Further, let us consider the following simple linear unweighted model:

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + e_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, n_i, \tag{3}$$

or in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{4}$$

where the definitions of all the variables are the same as in the previous case. Note that is a particular case of model (1) or (2) respectively, where random effects do not exist and the error variances are equal (heterocedasticity).

For the sake of brevity we skip developed formulas for model (3) or (4), because it is very easy to find maximum likelihood estimators of this model in the literature. We use the model in matrix form (2) to obtain the estimates of the parameters.

2.1 ML estimation in a mixed lineal regression model

ML estimates of $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\sigma}^t)^t = (\boldsymbol{\beta}^t, \sigma_0^2, \sigma_1^2)^t$ can be obtained by maximizing the log-likelihood function

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

with the Fisher-scoring algorithm

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{F}(\boldsymbol{\theta}^{(k)})^{-1}\boldsymbol{S}(\boldsymbol{\theta}^{(k)}),$$

where $\boldsymbol{S}(\boldsymbol{\theta})$ and $\boldsymbol{F}(\boldsymbol{\theta})$ are the $(p+2)\times 1$ vector of scores and the $(p+2)\times(p+2)$ Fisher information matrix, respectively. The block elements of $\boldsymbol{S}(\boldsymbol{\theta})$ and $\boldsymbol{F}(\boldsymbol{\theta})$ are

$$S_{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{I} \boldsymbol{X}_i{}^t \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),$$

$$\begin{aligned} S_{\sigma_1^2} &= -\frac{1}{2}\sum_{i=1}^{I} \operatorname{tr}\left\{\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\right\} \\ &\quad + \frac{1}{2}\sum_{i=1}^{I} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^t \boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \\ S_{\sigma_0^2} &= -\frac{1}{2}\sum_{i=1}^{I} \operatorname{tr}\left\{\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\right\} \\ &\quad + \frac{1}{2}\sum_{i=1}^{I} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^t \boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \end{aligned}$$

$$F_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^{I} \boldsymbol{X}_i^t\boldsymbol{V}_i^{-1}\boldsymbol{X}_i, \quad F_{\sigma_1^2\sigma_1^2} = \frac{1}{2}\sum_{i=1}^{I} \operatorname{tr}\left\{\left(\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\right)^2\right\},$$

$$F_{\sigma_1^2\sigma_0^2} = \frac{1}{2}\sum_{i=1}^{I} \operatorname{tr}\left\{\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\right\},$$

$$F_{\sigma_0^2\sigma_0^2} = \frac{1}{2}\sum_{i=1}^{I} \operatorname{tr}\left\{\left(\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\right)^2\right\}.$$

2.2 REML estimation in a mixed lineal regression model

The REML estimation method reduces the bias of the variance components that appear with the ML method. This method consists of estimating by one side the variance components and by the other side the fixed effects [13–15]. REML estimates of $\boldsymbol{\sigma} = (\sigma_0^2, \sigma_1^2)^t$ can be obtained by maximizing the log-likelihood function

$$\ell(\boldsymbol{\sigma}) = -\frac{1}{2}(n-p)\log 2\pi - \frac{1}{2}\log|\boldsymbol{K}^t\boldsymbol{V}\boldsymbol{K}| - \frac{1}{2}\,\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y},$$

where $\boldsymbol{K} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}$ and $\boldsymbol{P} = \boldsymbol{K}(\boldsymbol{K}^t\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{K}^t$. The updating equation of the Fisher-scoring algorithm is

$$\boldsymbol{\sigma}^{(k+1)} = \boldsymbol{\sigma}^{(k)} + \boldsymbol{F}(\boldsymbol{\sigma}^{(k)})^{-1}\boldsymbol{S}(\boldsymbol{\sigma}^{(k)}).$$

In this case the block elements of $\boldsymbol{S}(\boldsymbol{\sigma})$ and $\boldsymbol{F}(\boldsymbol{\sigma})$ are

$$S_{\sigma_1^2} = -\frac{1}{2}\,\mathrm{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\} + \frac{1}{2}\boldsymbol{y}^t\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\boldsymbol{P}\boldsymbol{y}\,,$$

$$S_{\sigma_0^2} = -\frac{1}{2}\,\mathrm{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\} + \frac{1}{2}\boldsymbol{y}^t\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\boldsymbol{P}\boldsymbol{y}\,,$$

$$F_{\sigma_0^2\sigma_0^2} = \frac{1}{2}\,\mathrm{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\},\,,$$

$$F_{\sigma_0^2\sigma_1^2} = \frac{1}{2}\,\mathrm{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\}\,,$$

$$F_{\sigma_1^2\sigma_1^2} = \frac{1}{2}\,\mathrm{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\}\,.$$

3 Simulation experiments

In this section we present three simulation experiments. The first simulation experiment is designed to compare the impact of the presence or absence of weights when the model (4) is fitted in two types of the target variables.

The second simulation experiment is designed to study the behavior of the model (2) under the ML and REML fitting methods.

The third simulation experiment is designed to study the robustness of the estimation methods (ML and REML) under the model (2); we repeat the second simulation experiment by changing the error normal distributions.

Numerical results of the simulation study are available via the web from: http://helmantica.umh.es/wessex/data2013/

3.1 Experiment 1

The scope of this simulation experiment is to investigate the impact of the presence or absence of weights when the model (4) is fitted on two types of simulated target variables. One of them, (Y0), is simulated as a linear regression model with a fixed effect, the other one, (Y1), is simulated as a mixed regression model (a fixed effect and a random effect).

Each data set is generated as follows. For $i = 1, \ldots, I$, $j = 1, \ldots, n_i$:

- Explanatory variable: $x_{ij} = (b_i - a_i)U_{ij} + a_i$ with $U_{ij} = \frac{j}{n_i+1}$. $a_i = 1$, $b_i = 1 + \frac{1}{I}(I + i)$.
- Weights: $w_{ij} = 1/x_{ij}^{\ell}$, $\ell = 0, 1/2, 1, 2$, (4 possibilities).
- Random effects and errors: $u_i \sim N(0, \sigma_1^2 = 1)$, $e_{ij} \sim N(0, \sigma_0^2 = 1)$.
- Two target variables: calculate

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + w_{ij}^{-1/2} e_{ij}, \quad \text{with } \beta_0 = \beta_1 = 1, \tag{Y0}$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + w_{ij}^{-1/2} e_{ij}, \quad \text{with } \beta_0 = \beta_1 = 1. \tag{Y1}$$

The simulation experiment follows the steps:

1. Repeat $K = 10^6$ times ($k = 1, \ldots, K$)
   1.1. Generate a data set of size $n = \sum_{i=1}^{I} n_i$.
   1.2. Calculate $\tau_{(k)} \in \{\widehat{\boldsymbol{\beta}}_{(k)}, \widehat{\sigma}^2_{0,(k)}\}$ and fitted values by using the ML method with the unweighted simple linear model (4).
2. Calculate, for every $\tau \in \{\boldsymbol{\beta}, \sigma_0^2\}$ and fitted values, the scaled EMSEs and BIASes

$$\text{EMSE}(\widehat{\tau}) = \frac{10^6}{K} \sum_{k=1}^{K} (\widehat{\tau}_{(k)} - \tau)^2, \qquad \text{BIAS}(\widehat{\tau}) = \frac{10^6}{K} \sum_{k=1}^{K} (\widehat{\tau}_{(k)} - \tau).$$

$$\text{EMSE}(\widehat{\mu}) = \frac{10^6}{K} \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\widehat{y}_{ij(k)} - y_{ij})^2,$$

$$\text{BIAS}(\widehat{\mu}) = \frac{10^6}{K} \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\widehat{y}_{ij(k)} - y_{ij}).$$

The simulations are carried out for the seven combinations of sizes presented in Table 1.

In the simulation experiment we focus our attention on two performance measures: the empirical mean squared error (EMSE) and the empirical bias (BIAS). Figures 1 and 2 plot the EMSE and BIAS results respectively. These figures are divided into four parts, one for each parameter. Each part is again divided into four sections, the left one for homocedasticity ($\ell = 0$) and the other three for heterocedasticity ($\ell = 1/2, 1, 2$). To better interpret the figures, total sample size n and total number of levels I are plotted (see the lower and upper horizontal axes).

Table 1: Groups of datasets sizes.

g	1	2	3	4	5	6	7
$I^{(g)}$	5	7	10	20	30	50	75
n_i	100	100	100	100	100	100	100
n	500	700	1000	2000	3000	5000	7500

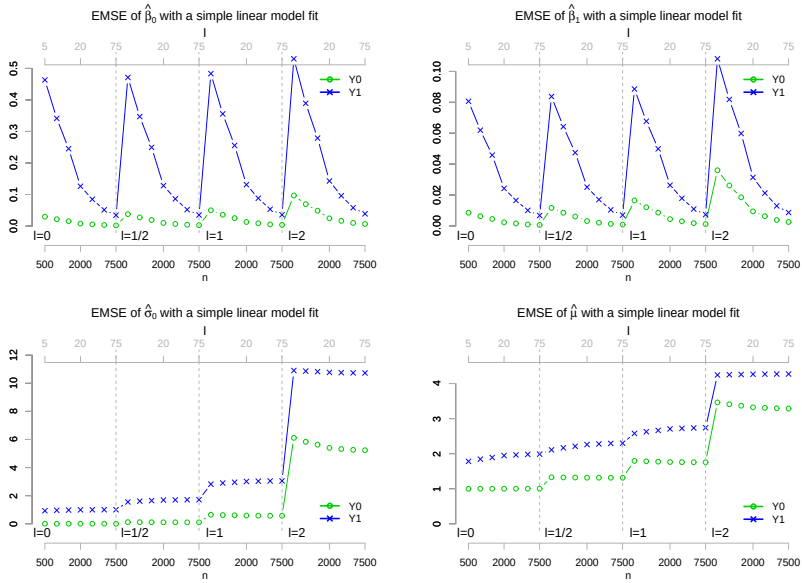

Figure 1: EMSE of $\widehat{\boldsymbol{\beta}}$ (top), $\widehat{\sigma}_0^2$ (bottom-left) and $\widehat{\mu}$ (bottom-right), for $\ell = 0, 1/2, 1, 2$.

Regarding EMSE, Figure 1 shows two different conclusions. First of all is that the model fitt is better when the target variable simulated only has fixed effects. This is very logical, since if the model proposed is (4), when it fits a dataset with a random factor, the EMSE must grow. The second conclusion is about the heterocedasticity. The presence of heterocedasticity ($\ell = 1/2, 1, 2$) damages the estimation of all the parameters in both target variables when an unweighted model is used to fit. Regarding BIAS, Figure 2 shows some interesting patterns. If the target is to estimate $\boldsymbol{\beta}$, model (4) in Y0 variable (green line) presents lower biases than in Y1 variable (blue line). The presence of heterocedasticity ($\ell = 1/2, 1, 2$) causes bias in the estimation of σ_0^2. However, if we look at the goodness of fit, both target variables are near unbiasedness. Of course, Y0 is better than in the other.

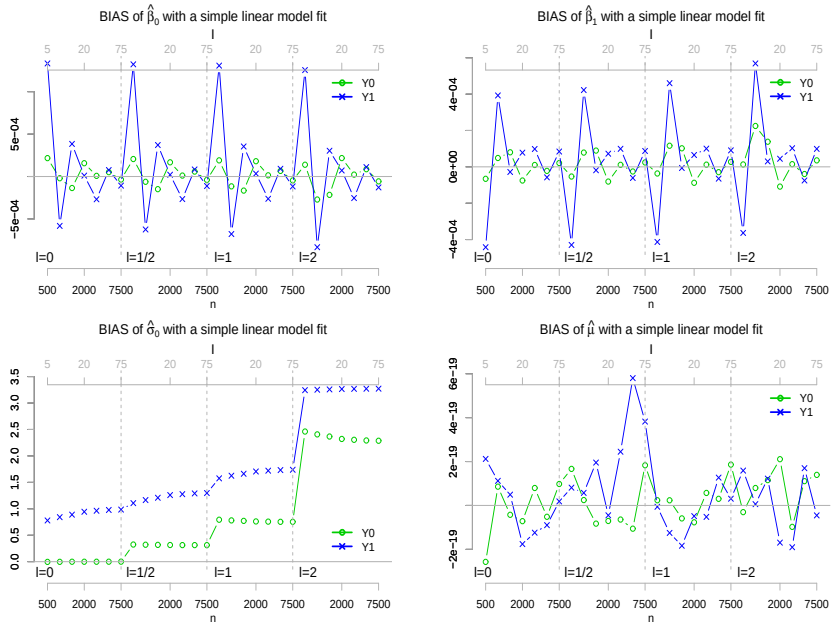

Figure 2: BIAS of $\widehat{\boldsymbol{\beta}}$ (top), $\widehat{\sigma}_0^2$ (bottom-left) and $\widehat{\mu}$ (bottom-right), for $\ell = 0, 1/2, 1, 2$.

3.2 Experiment 2

This simulation experiment is designed to study the behavior of the model (2) under the ML and REML fitting methods. These methods are tested on two types of simulated target variables. Target variables (Y0 and Y1), explanatory variables and weights are generated in the same way as in the first simulation experiment. The steps of the simulation experiment are also the same except for the step of the model fit, which reads as follows:

1.2. Calculate $\tau_{(k)} \in \{\widehat{\boldsymbol{\beta}}_{(k)}, \widehat{\sigma}^2_{1,(k)}, \widehat{\sigma}^2_{0,(k)}\}$ and fitted values by using the REML and ML methods with the mixed linear model (2).

For more details see subsection 3.1. The simulations are carried out for the seven combinations of sizes presented in Table 1.

In this simulation experiment we focus our attention on both estimation methods. Figures 3 and 4 plot the EMSE and BIAS results respectively for the REML methods. The figures have the same structure as in the previous subsection. Figures like this for ML are ignored because the plot lines and the values are almost identical. Figure 5 plots the EMSE and BIAS results for the difference between REML and ML methods.

In order for the reader to see the values of the differences between the estimation methods, this graph is plotted instead of the other two. Superindexes ML are REML are simply used to denote that the EMSE or BIAS has been calculated by using ML or REML estimates of the model parameters. Regarding REML,

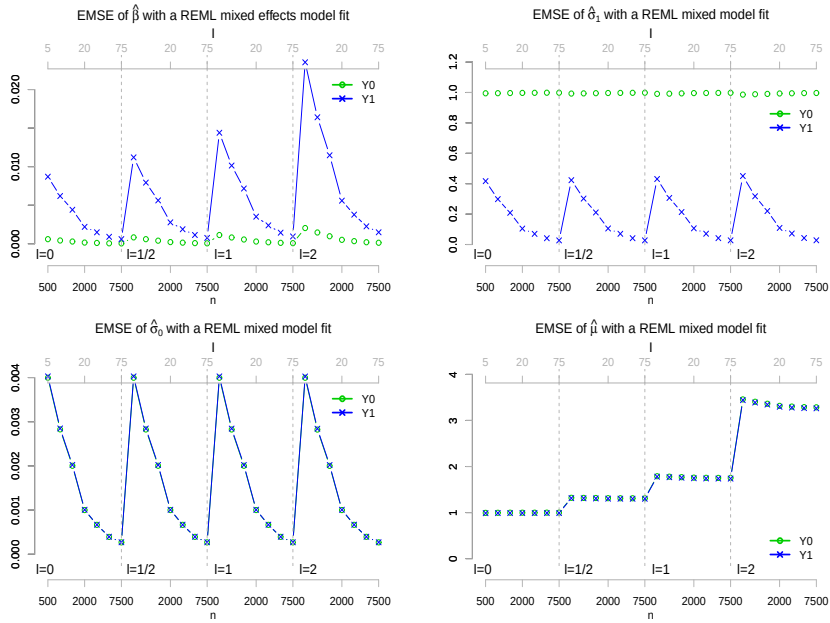

Figure 3: EMSE of $\widehat{\beta}$ (top-left), $\widehat{\sigma}_1^2$ (top-right), $\widehat{\sigma}_0^2$ (bottom-left) and $\widehat{\mu}$ (bottom-right), for $\ell = 0, 1/2, 1, 2$.

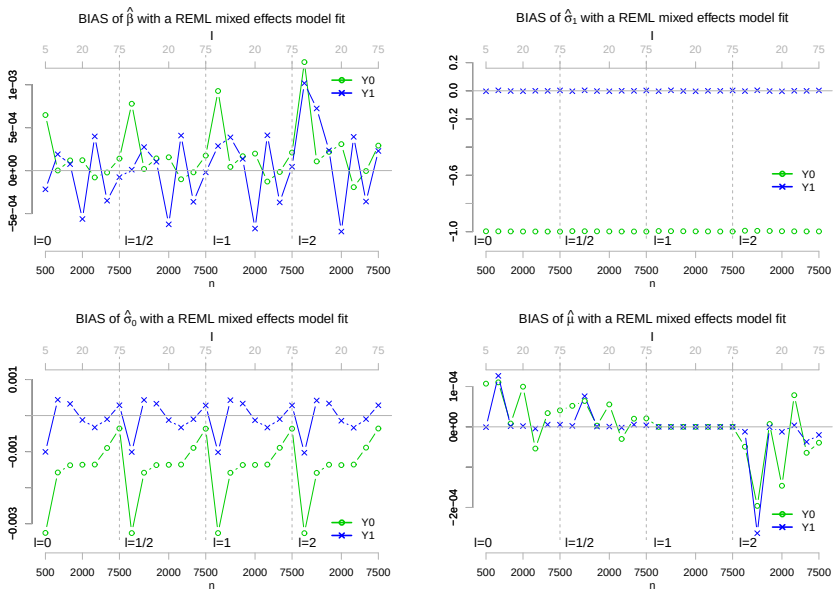

Figure 4: BIAS of $\widehat{\beta}$ (top-left), $\widehat{\sigma}_1^2$ (top-right), $\widehat{\sigma}_0^2$ (bottom-left) and $\widehat{\mu}$ (bottom-right), for $\ell = 0, 1/2, 1, 2$.

Figures 3 and 4 shows two different conclusions. First of all is that the model fitted is a bit better when the target variable simulated have mixed effects. The

EMSE is only better for the random effect variance and reduces when the dataset is larger. Moreover, the prediction is worse in the presence of heterocedasticity, but prediction is better for (Y1) target as expected (numerical result can be seen on the web).

The second conclusion is about the BIAS shown in Figure 4. Only for the (Y1) target variable, is the REML method is unbiased. The presence or absence of heterocedasticity does not affect the bias at all.

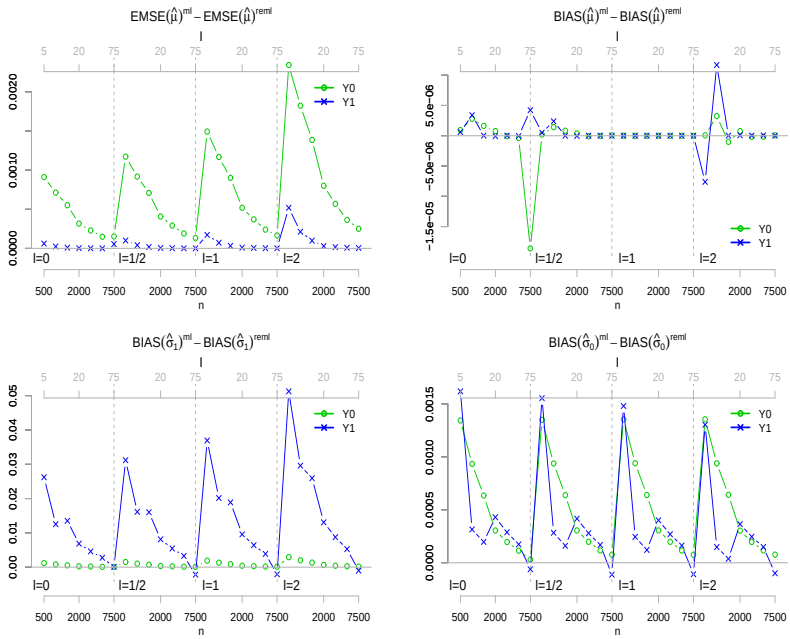

Figure 5: EMSE$^{\text{ml}}$ − EMSE$^{\text{reml}}$ for μ (top-left), BIAS$^{\text{ml}}$ − BIAS$^{\text{reml}}$ for μ (top-right), BIAS$^{\text{ml}}$ − BIAS$^{\text{reml}}$ for σ_1^2 (bottom-left) and BIAS$^{\text{ml}}$ − BIAS$^{\text{reml}}$ for σ_0^2 (bottom-right), for $\ell = 0, 1/2, 1, 2$.

Figure 5 always shows positive numbers, so then ML estimators are worse than REML ones. This is especially relevant underthe (Y1) target variable, as can be seen in the EMSE of the predictions.

3.3 Experiment 3

This simulation experiment is designed to study the robustness of the model (2) under the ML and REML fitting methods. We repeat the second simulation experiment by changing the error normal distributions to Gamma and Weibull distributions. This give us information about the adequation of this model and these methods with respect to deviations from the assumption of normality of the errors. Gamma a Weibull distributions are conveniently parametrized to have the same means and variances as in the normal case (Ga(1, 1), We(1, 1)). Note that Gamma and Weibull distributions have positives supports. This is very important

when the variable is a currency. The Weibull distribution is sometimes used to model claims in reinsurances.

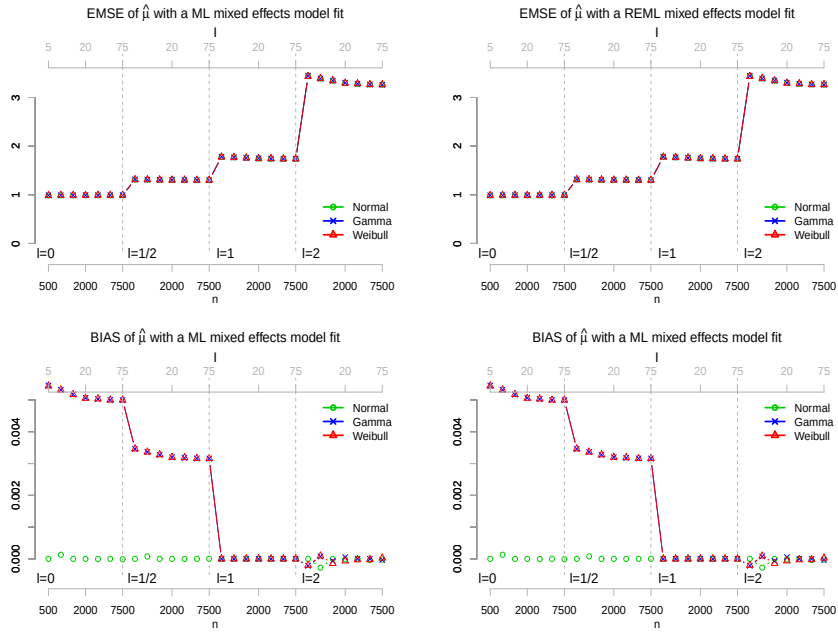

Figure 6: EMSE$(\widehat{\mu})^{\text{ml}}$ and EMSE$(\widehat{\mu})^{\text{reml}}$ (top) and BIAS$(\widehat{\mu})^{\text{ml}}$ and BIAS$(\widehat{\mu})^{\text{reml}}$ (bottom), $\ell = 0, 1/2, 1, 2$ for Normal, Gamma, and Weibull cases.

The obtained results for the predictions are presented in Figure 6. The first fact that we observe is that the results are practically identical for all the cases between ML (left) and REML (right). Then the following comments are applicable to both methods. In the presence of heterocedasticity, EMSE grows while the BIAS decreases to 0. Only under homocedasticity or near it, is the normal case is unbiased. Concerning heterocedasticity, the BIAS seems to be unaffected by deviations from normality.

4 Conclusions

In this paper, two models are introduced and simulation studies carried out to investigate when it is worthwhile to predict bank credit risk.

The first simulation experiment is designed to prove a linear regression model. The second simulation is designed to study the behavior of the linear mixed model under two estimation methods. The third simulation experiment is designed to study the robustness of the estimation methods by changing the error normal distributions.

From the obtained results we conclude recommending the use of REML estimates in a mixed model and being careful with the heterocedasticity in the

linear fixed model. We remark that deviations from the assumption of normality in the random errors do not change the predictions at all.

It is important to realize that when the number of levels of the random factor increases, the number of parameters for the estimate is the same. Only one variance parameter is estimated. A fixed effects model is estimated with as many parameters as the number of levels of the factor minus one. In datasets of home equity loans, it is very common to find factors with a larger number of levels, and then it is necessary to approach a mixed linear model.

Acknowledgements

The authors are grateful to the University Miguel Hernández of Elche for the financial support under the grant "Proyectos de Investigación en Humanidades y Ciencias Sociales". That has made the research for this article possible.

References

[1] Fisher, R., The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, pp. 179–188, 1936.
[2] Durand, D., *Risk Elements in Consumer Instalment Financing*. National Bureau of Economic Research: Massachusetts, 1941.
[3] Hand, D. & Henley, W.E., Statistical classification methods in consumer credit scoring: a review. *Royal Statistical Society*, **160(3)**, pp. 523–541, 1997.
[4] Srinivasan, V. & Kim, Y.H., Credit granting: a comparative analysis of classification procedures. *The Journal of Finance*, **XLII(3)**, pp. 665–683, 1987.
[5] Tam, K. & Kiang, M., Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, **38**, pp. 926–947, 1992.
[6] Desai, V., Crook, J. & Overstreet, G., A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, **95**, pp. 24–37, 1996.
[7] Yobas, Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, **11**, pp. 111–125, 2000.
[8] Yatchew, A., Non parametric regression techniques in economics. *Journal of Economic Literature*, **16**, pp. 669–721, 1998.
[9] Boj, E., Claramunt, M.M., Grané, A. & Fortiana, J., Projection error term in gower's interpolation. *Journal of Statistical Planning and Inference*, **139**, pp. 1867–1878, 2009.
[10] Minsky, H., The financial instability hypothesis. *The Jerome Levy Economics Institute*, **74**, 1992.
[11] Thomas, L.C., A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16**, pp. 149–172, 2000.

[12] Artís, M., Guillén, M. & Martínez, J.M., A model for credit scoring: an application of discriminant analysis. *QÜESTIIÓ*, **18(3)**, pp. 385–395, 1994.
[13] Searle, S., Casella, G. & McCullogh, C., *Variance components*. John Wiley and Sons, Inc.: New-York, 1982.
[14] Bartlett, M.S., Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **160**, pp. 1405–1419, 1937.
[15] Njuho, P. & Milliken, G., Analysis of linear models with one factor having both fixed and random effects. *Communications in Statistics - Theory and Methods*, **34**, pp. 1979–1989, 2005.